




# Introduction à l'analyse méta-omics

Claire Hoede

[https://web-genobioinfo.toulouse.inrae.fr/~formation/M2\\_bioinfo/2024/](https://web-genobioinfo.toulouse.inrae.fr/~formation/M2_bioinfo/2024/)



- Présentation de la PF Bioinfo
- Tour de table

1. Introduction
  2. Applications
  3. Analyses méta-omiques
    - 3.1. La méta-génétique
    - 3.2. La méta-génomique
  4. Analyses exploratoires
  5. Impact carbone du calcul
- 

# 1. Introduction

## 2. Applications

## 3. Analyses méta-omiques

3.1. La méta-génétique

3.2. La méta-génomique

## 4. Analyses exploratoires

## 5. Impact carbone du calcul





# 1. Introduction

Quelques définitions

**Metagénomique :**



# 1. Introduction

## Quelques définitions

**Metagénomique** : utilisation des techniques de génomique modernes pour l'étude des communautés microbiennes directement dans leur environnement naturel contournant le besoin d'isolation et de culture des espèces individuelles. (Chen & Pachter, 2005)



# 1. Introduction

## Quelques définitions

**Metagénomique** : utilisation des techniques de génomique moderne pour à l'étude des communautés microbiennes directement dans leur environnement naturel contournant le besoin d'isolation et de culture des espèces individuelles. (Chen & Pachter, 2005)

**Micro-organismes** :



# 1. Introduction

## Quelques définitions

**Metagénomique** : utilisation des techniques de génomique modernes pour l'étude des communautés microbiennes directement dans leur environnement naturel contournant le besoin d'isolation et de culture des espèces individuelles. (Chen & Pachter, 2005)

**Micro-organismes** : ensemble des organismes microscopiques (bactéries, archées, virus, champignons et algues microscopiques).  
Ils jouent un rôle vital dans le fonctionnement général de la biosphère en participant par exemple au cycle du carbone ou de l'azote.  
Ils synthétisent l'oxygène de notre atmosphère, préservent les animaux (dont nous) de très nombreuses maladies ...  
Leur diversité spécifique et fonctionnelle, les mécanismes régissant leur dispersion ainsi que leur histoire évolutive demeurent encore mal connus.

# 1. Introduction

Quelques définitions

**Le microbiote :**



# 1. Introduction

## Quelques définitions

**Le microbiote** : ensemble des micro-organismes vivants au sein d'un environnement spécifique (incluant les non-cultivables)

**Microbiome** :

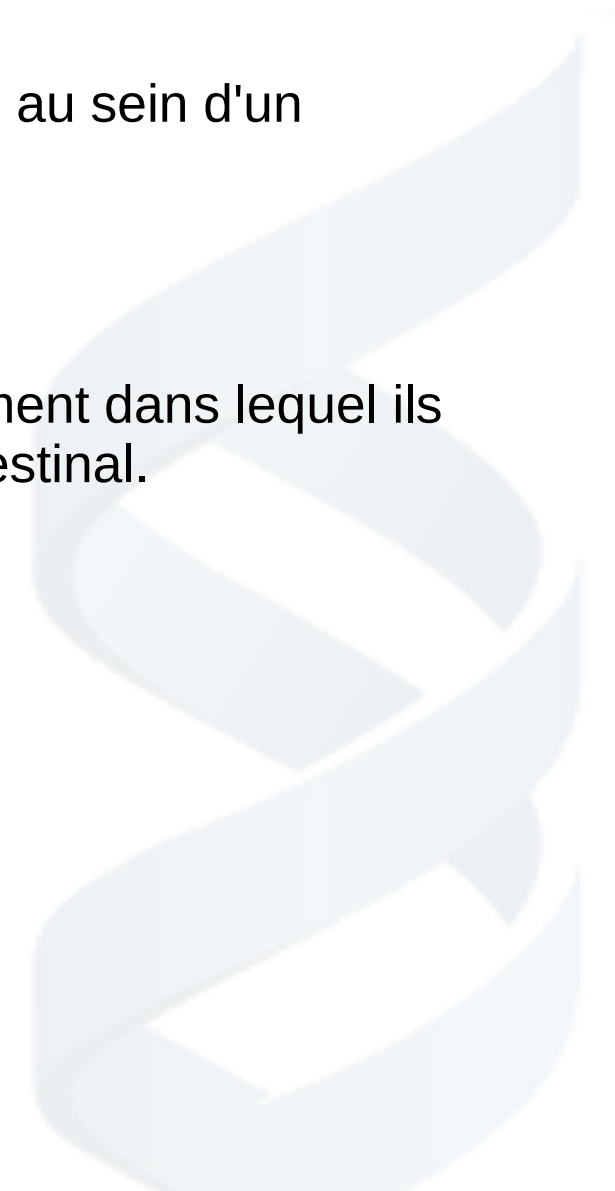


# 1. Introduction

## Quelques définitions

**Le microbiote** : ensemble des micro-organismes vivants au sein d'un environnement spécifique (incluant les non-cultivables)

**Microbiome** : ensemble du microbiote et de l'environnement dans lequel ils évoluent et interagissent. Par exemple le microbiome intestinal.



# 1. Introduction

## Pourquoi les espèces non cultivables ?

Une faible proportion de micro-organismes sont cultivables.

Mis à part quand ils induisent une pathologie, les micro-organismes peuvent être difficilement détectables.

La plupart des micro-organismes ne sont pas pathogènes (même ceux qui sont associés à l'homme).

Pour toutes ces raisons les techniques d'études ne nécessitant pas une culture cellulaires ont été développées ces dernières années (typiquement l'analyse de sequence marqueur tels que l'ARNr 16S, 18S, ITS et la metagenomique)



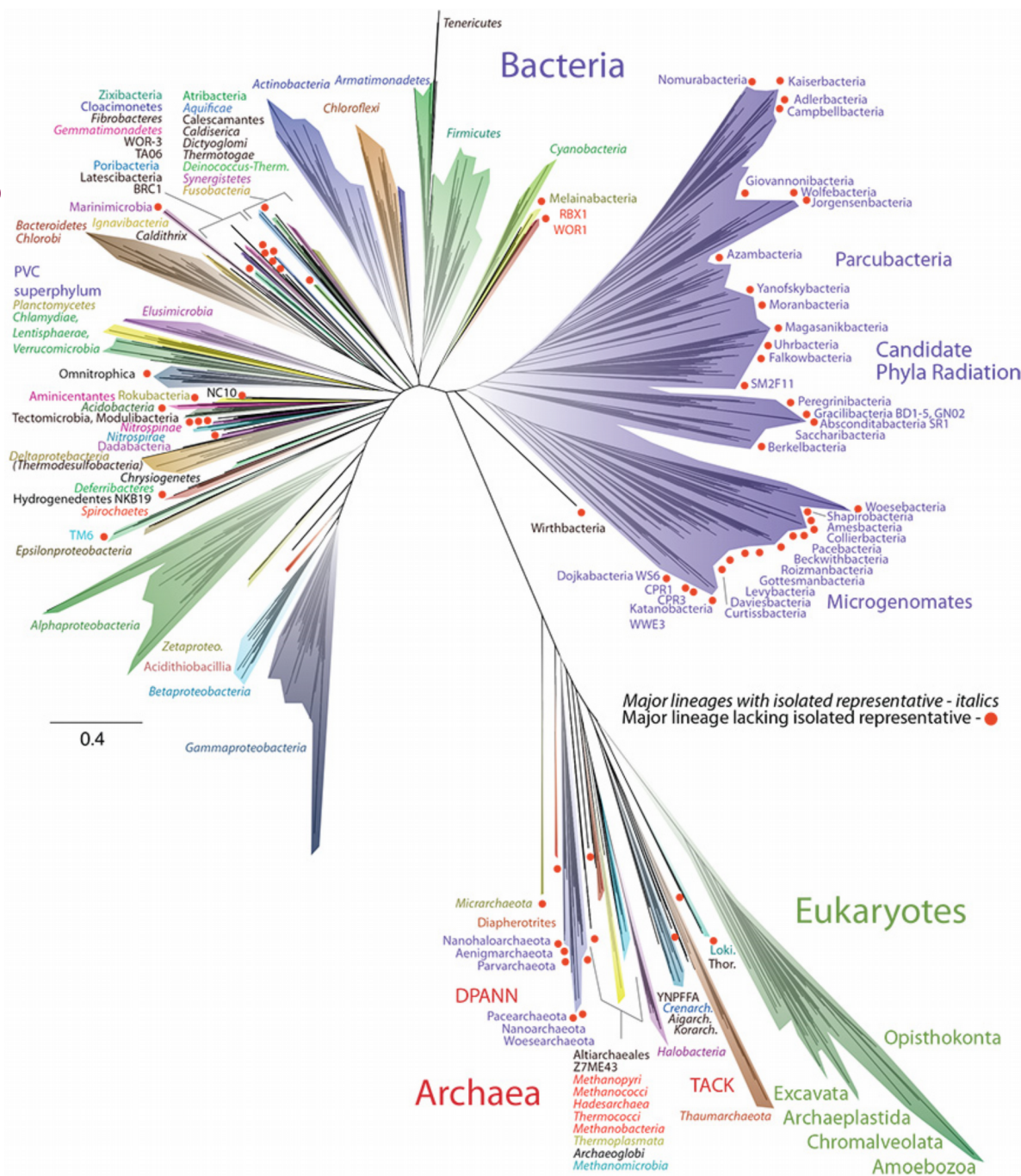
# 1. Introduction

## Espèces non cultivables ?

**A new view of the Tree of Life**  
(Hug et al. 2016)

Une étude de l'université de Berkeley montre que lorsqu'on inclut les données génomiques de 1000 organismes bactériens non cultivés issus d'échantillons métagénomiques, on a une vision complètement différente de l'arbre de la vie et de la diversité des 3 grands domaines du vivant.

Arbre construit à partir de 16 protéines ribosomales.

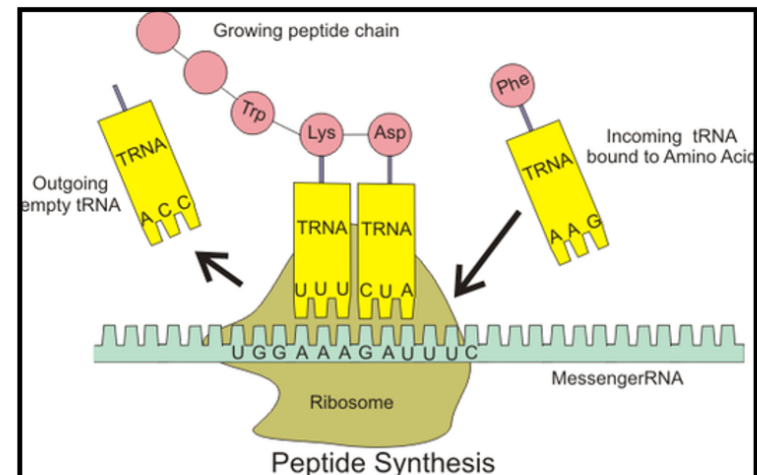
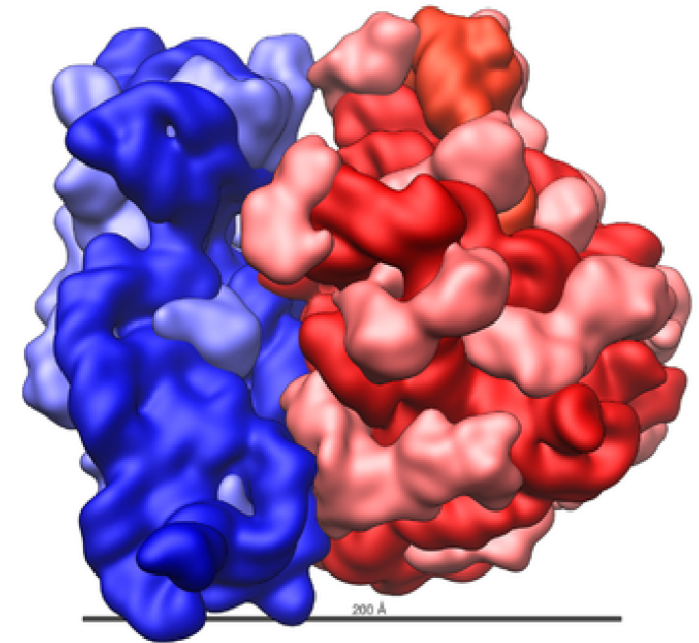


# 1. Introduction

## Ribosomes

- Sont composés de protéines et d'ARNr.
- Extrêmement conservés au cours de l'évolution.
- Présents dans les cellules eucaryotes et procaryotes.
- Synthétisent les protéines en décodant l'information contenue dans l'ARN messenger.
- L'ARN ribosomique portent l'activité catalytique, les protéines stabilisent la structure.
- Les ribosomes sont constitués de deux sous-unités, la petite « lit » l'ARN messenger et la grosse se charge de la polymérisation des acides aminés pour former la protéine correspondante.

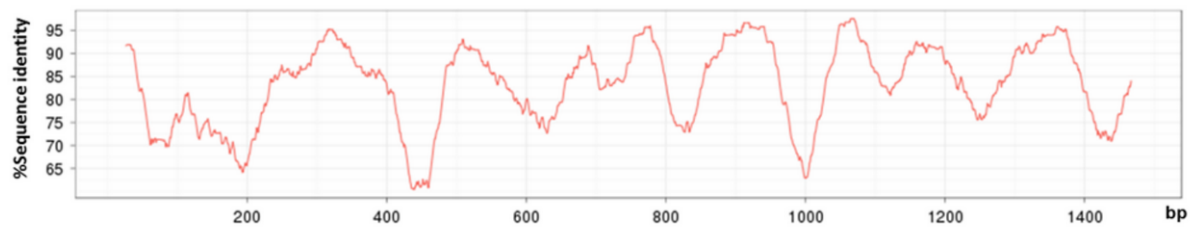
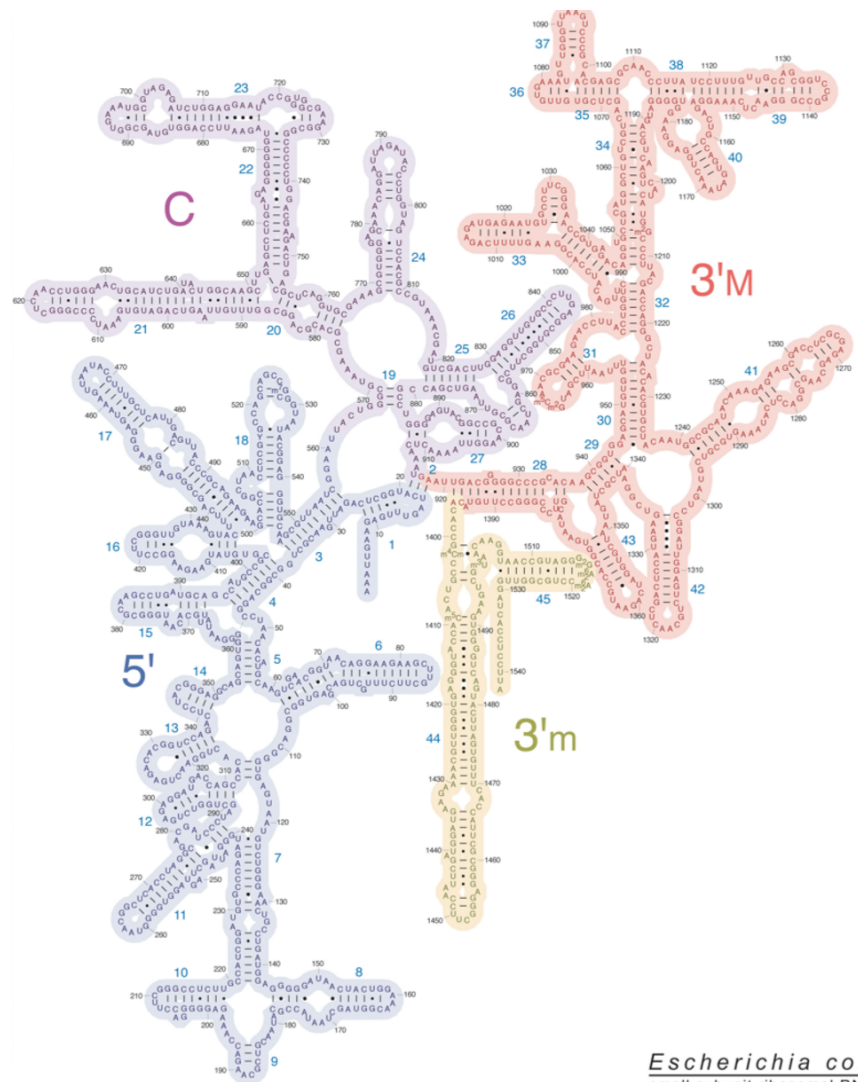
ribosome



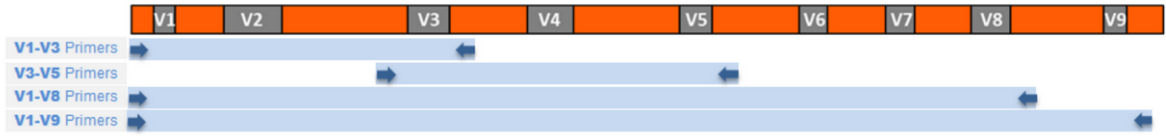
# 1. Introduction

## petite sous unité ribosomale

- Ubiquitaire (16S pour les bactéries, 18S pour les eucaryotes)
- non soumis au transfert horizontal
- fonction constante (traduction)
- base de données disponibles (SILVA)



**CONSERVED REGIONS:** unspecific applications  
**VARIABLE REGIONS:** group or species-specific applications



# 1. Introduction

Les questions

**Qui est là ?**

**Que peuvent-ils faire ?**

(Que font-ils ? ==>  
metatranscriptomique...)

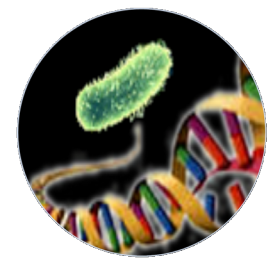
**Qu'est-ce que cela signifie ?**



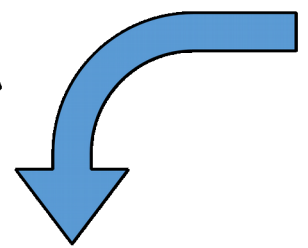


# 1. Introduction

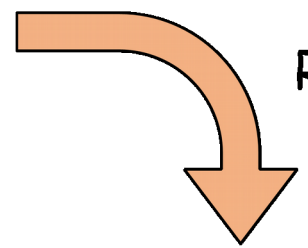
## Les questions



DNA



RNA



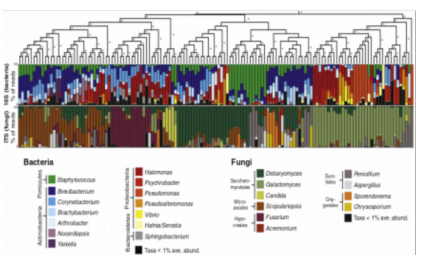
Metagenomics

Metatranscriptomics

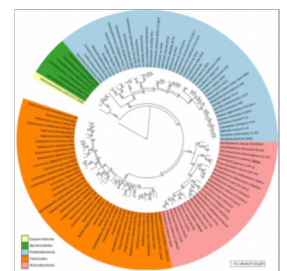
Amplicon sequencing

Shotgun sequencing

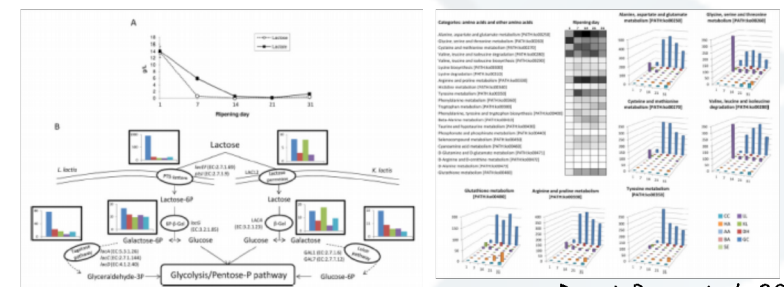
RNA sequencing



Wolfe et al., 2014



Almeida et al., 2014



Dugat-Bony et al., 2015

Who is here?

What can they do?

What are they doing?

Méta-génétique

Méta-génomique

# 1. Introduction

## Méta-génomique (la partie biologique)

### Extraction de l'ADN

### Amplification de la région ciblée (PCR)

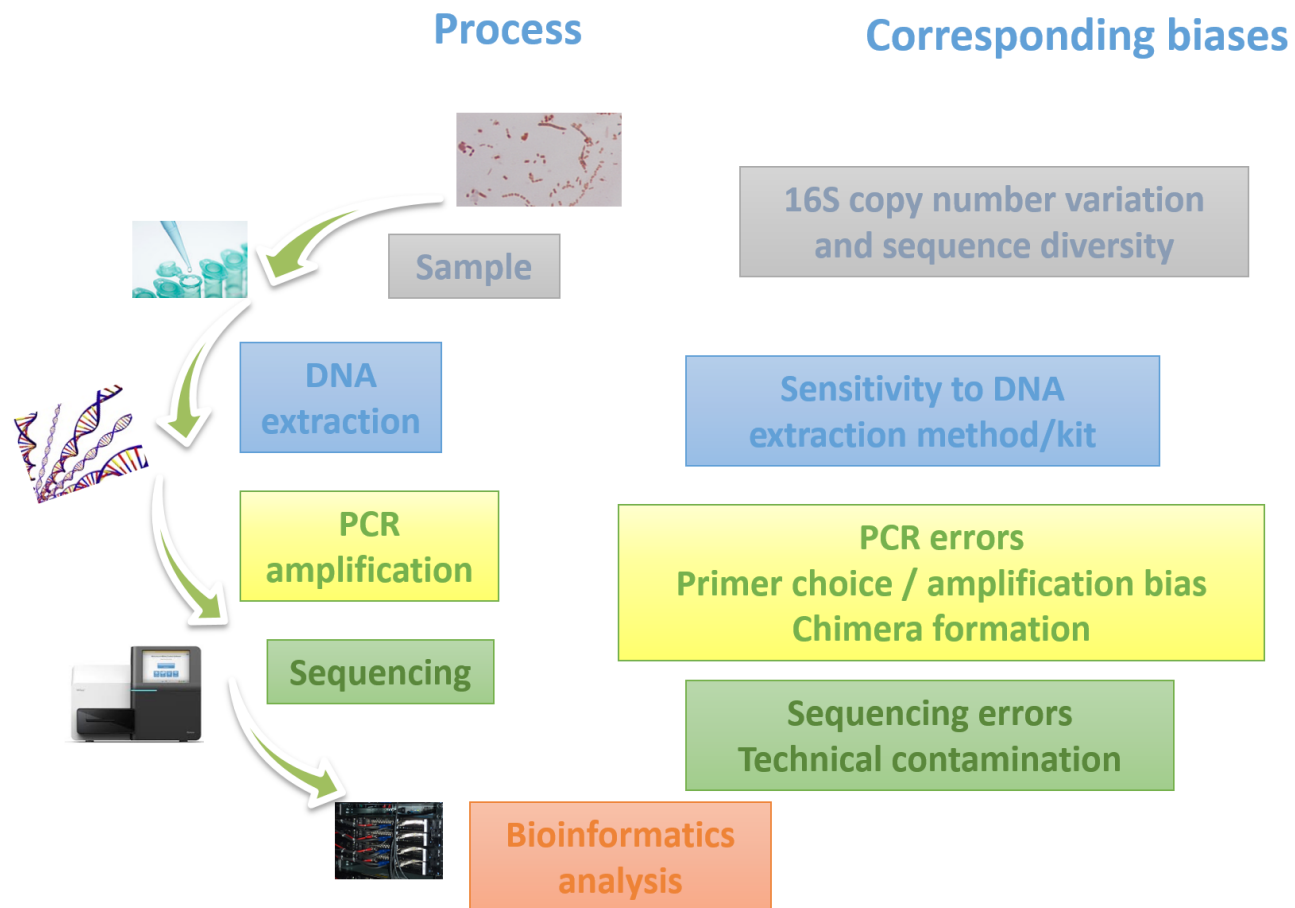
- amorce de PCR (primer) correspondant à l'extrémité de la région conservée autour de la région variable d'intérêt
- adaptateur pour le séquençage
- barcodes (petites séquences pour distinguer les échantillons multiplexés)



**Séquençage** généralement en MiSeq (2x250 pb, ~15 M paires de reads / run)  
On commence à voir des études qui séquencent le 16S complet ou le 16S-23S en longues lectures HiFi Pacbio ou ONT.

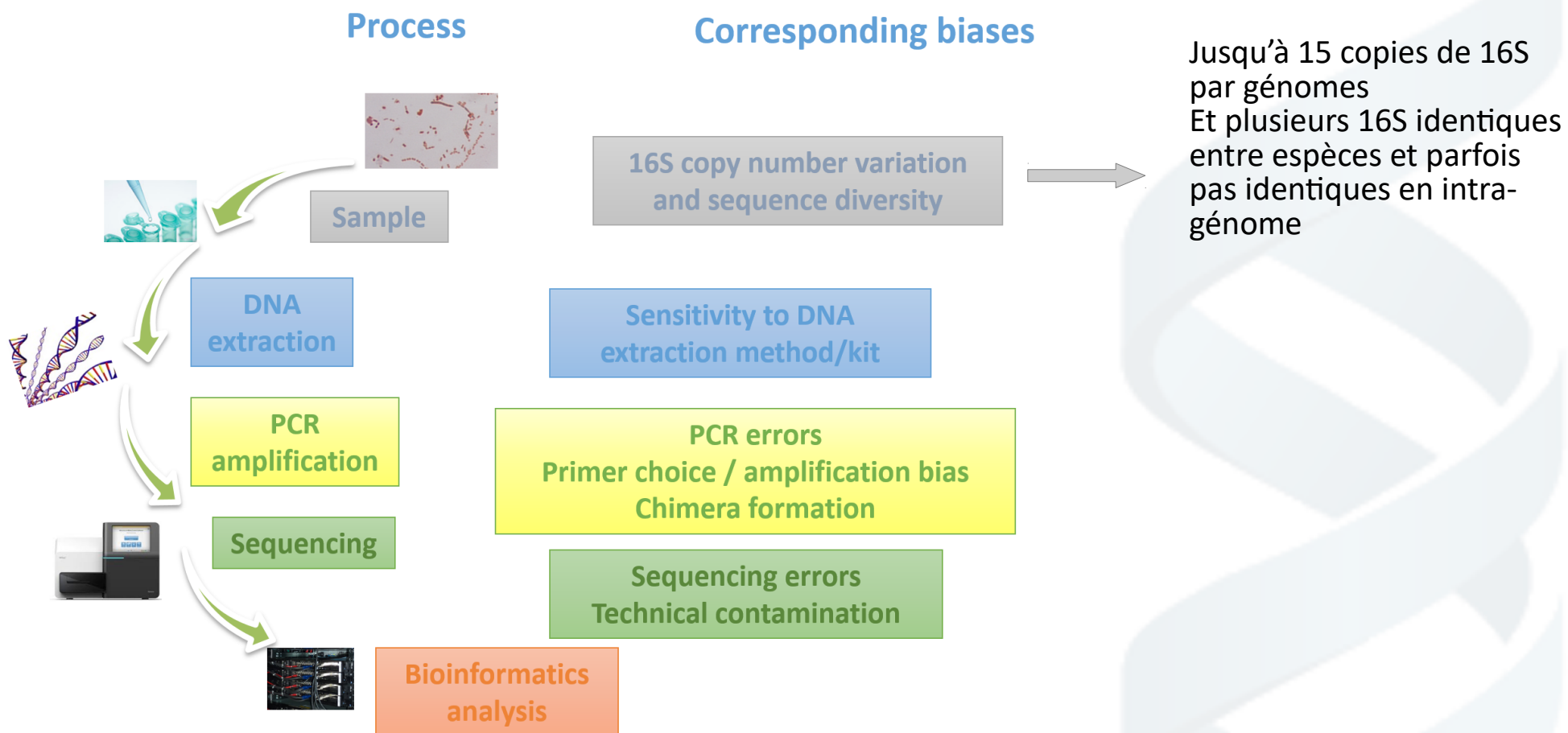
# 1. Introduction

## Méta-génétique (la partie biologique)



# 1. Introduction

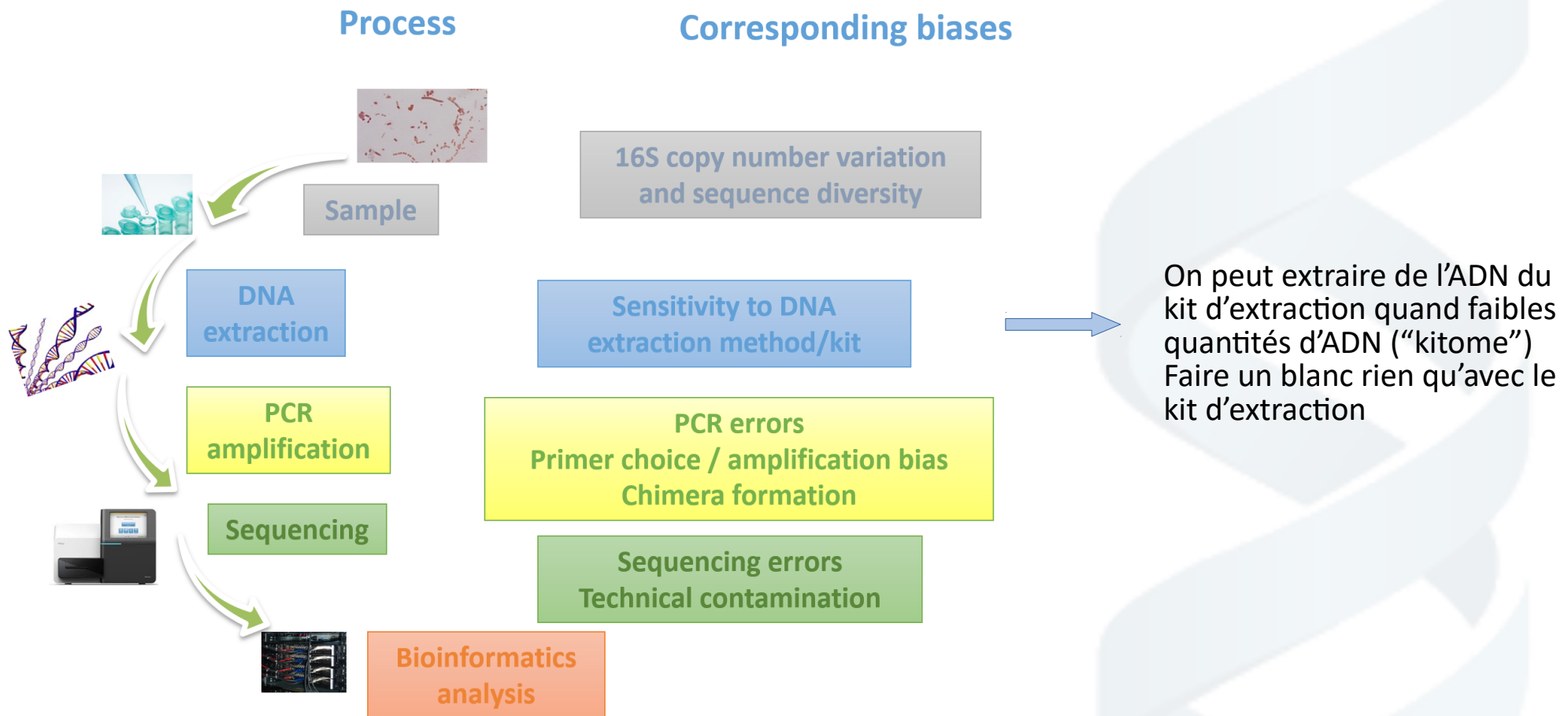
## Méta-génétique (la partie biologique)





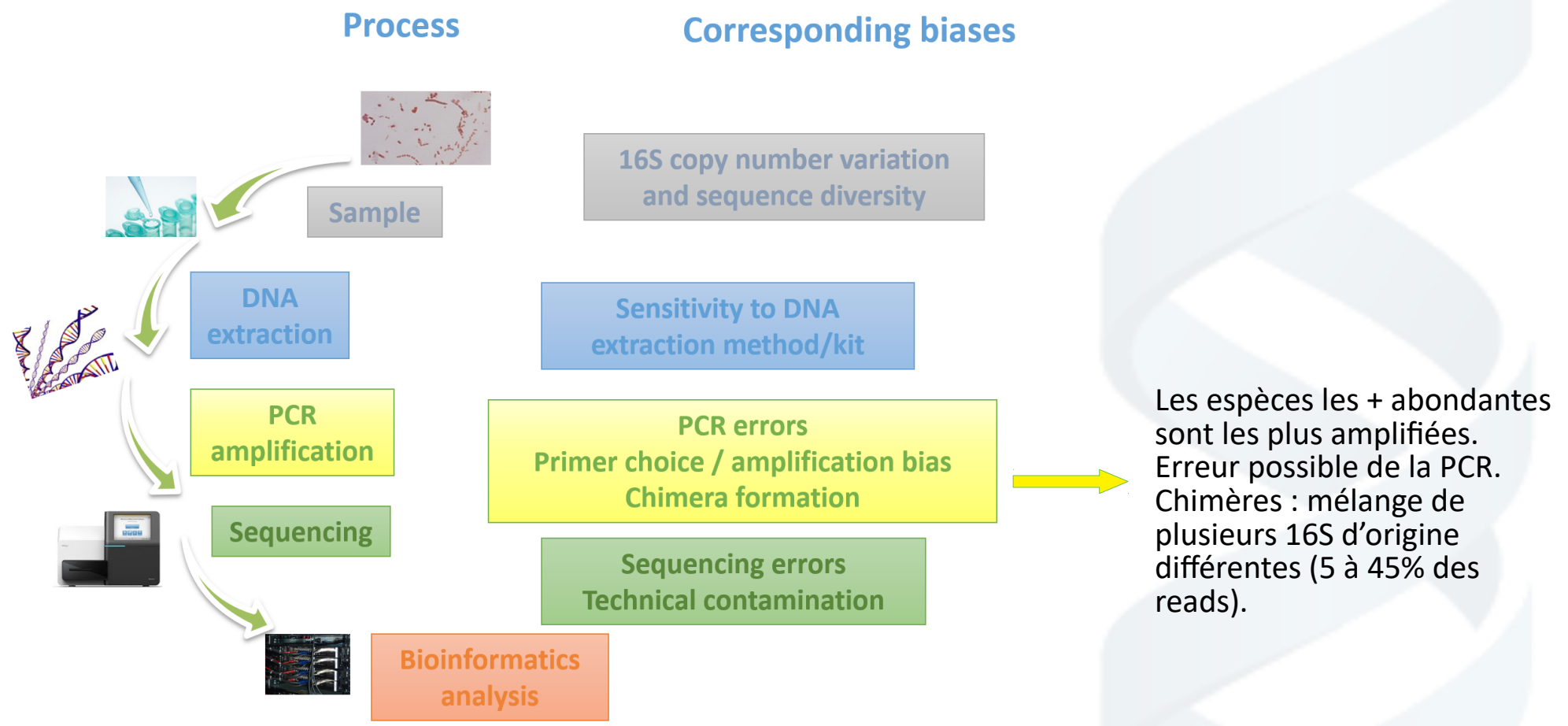
# 1. Introduction

## Méta-génétique (la partie biologique)



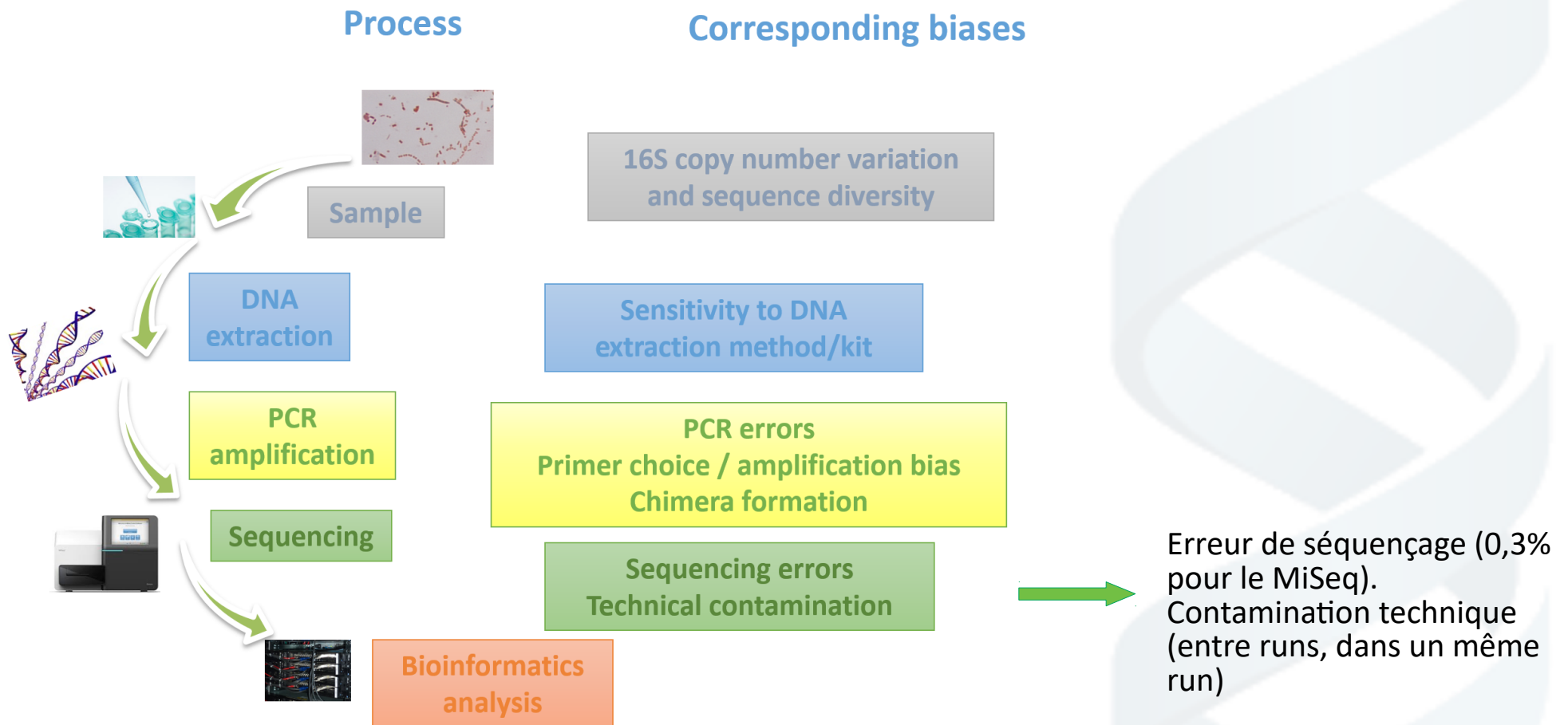
# 1. Introduction

## Méta-génétique (la partie biologique)



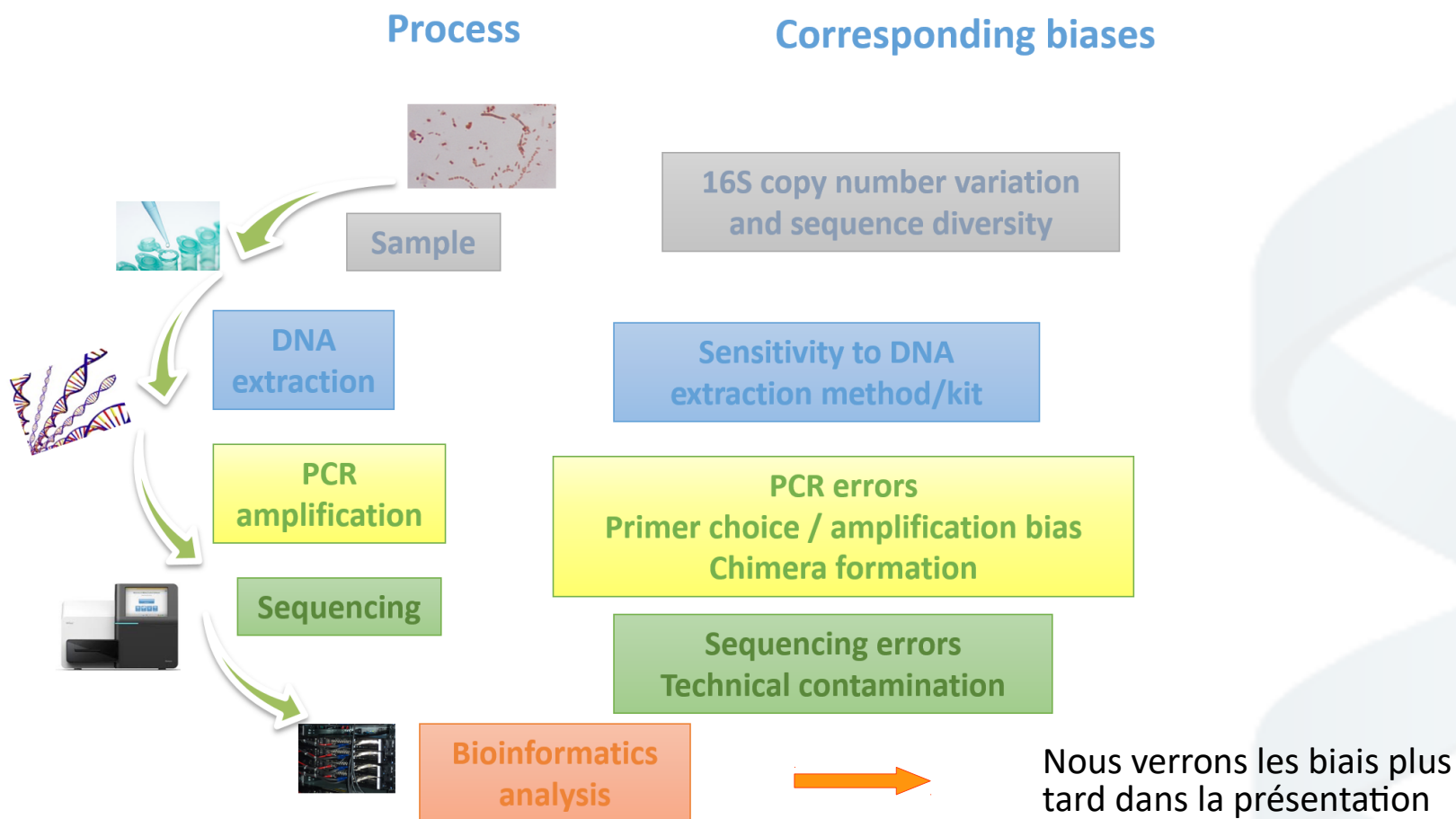
# 1. Introduction

## Méta-génétique (la partie biologique)



# 1. Introduction

## Méta-génétique (la partie biologique)

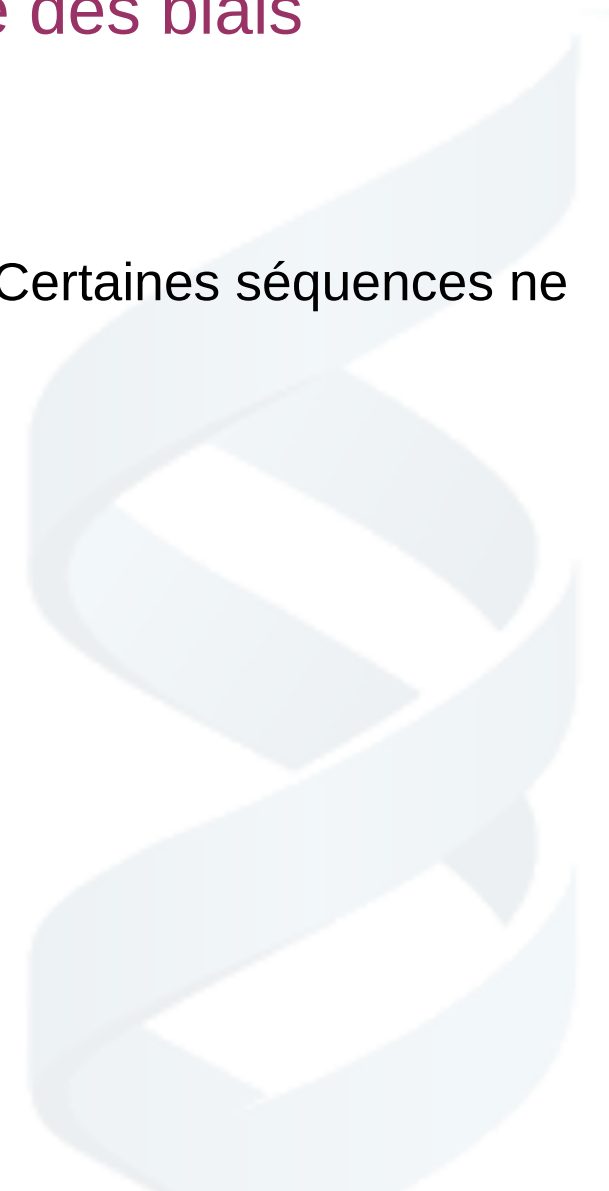


# 1. Introduction

## Méta-génomique (partie biologique) : résumé des biais

### **Biais biologiques :**

- nombre variable de copies de l'ARN 16S
- diversité intra-organisme
- la diversité des séquences diffère selon les clades. Certaines séquences ne varient pas.



# 1. Introduction

## Méta-génomique (partie biologique) : résumé des biais

### **Biais biologiques :**

- nombre variable de copies de l'ARN 16S
- diversité intra-organisme
- la diversité des séquences diffère selon les clades. Certaines séquences ne varient pas.

### **Biais techniques :**

- erreurs de PCR
- erreurs de séquençage
- biais d'amplification par PCR
- chimères
- méthodes d'extraction de l'ADN / kit
- contamination technique
- faible quantité d'ADN
- choix du séquenceur



# 1. Introduction

## Méta-génomique (partie biologique) : résumé des biais

### **Biais biologiques :**

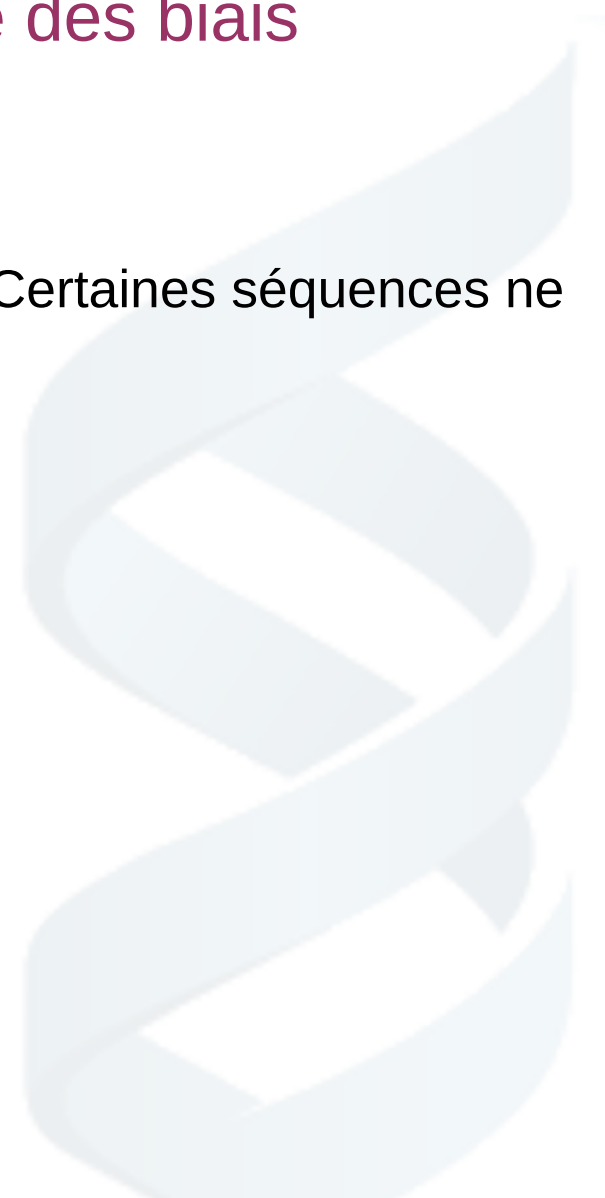
- nombre variable de copies de l'ARN 16S
- diversité intra-organisme
- la diversité des séquences diffère selon les clades. Certaines séquences ne varient pas.

### **Biais techniques :**

- erreurs de PCR
- erreurs de séquençage
- biais d'amplification par PCR
- chimères
- méthodes d'extraction de l'ADN / kit
- contamination technique
- faible quantité d'ADN
- choix du séquenceur

### **Biais humains :**

- contamination des échantillons
- choix de la région variable à amplifier
- choix des primers



# 1. Introduction

## Méta-génétique (partie biologique) : résumé des biais

### Biais biologiques :

- nombre variable de copies de l'ARN 16S
- diversité intra-organisme
- la diversité des séquences diffère selon les clades. Certaines séquences ne varient pas.

### Biais techniques :

- erreurs de PCR
- erreurs de séquençage
- biais d'amplification par PCR
- chimères
- méthodes d'extraction de l'ADN / kit
- contamination technique
- faible quantité d'ADN
- choix du séquenceur

### Biais humains :

- contamination des échantillons
- choix de la région variable à amplifier
- choix des primers



Biais sur la région variable choisie : donne pas les mêmes résultats surtout en terme d'abondance relative

On ne peut pas comparer les analyses faites sur des Vi différents

Si on veut comparer deux analyses il faut choisir les mêmes primers et les mêmes régions variables.

De plus la région choisie influence le nombre de chimères

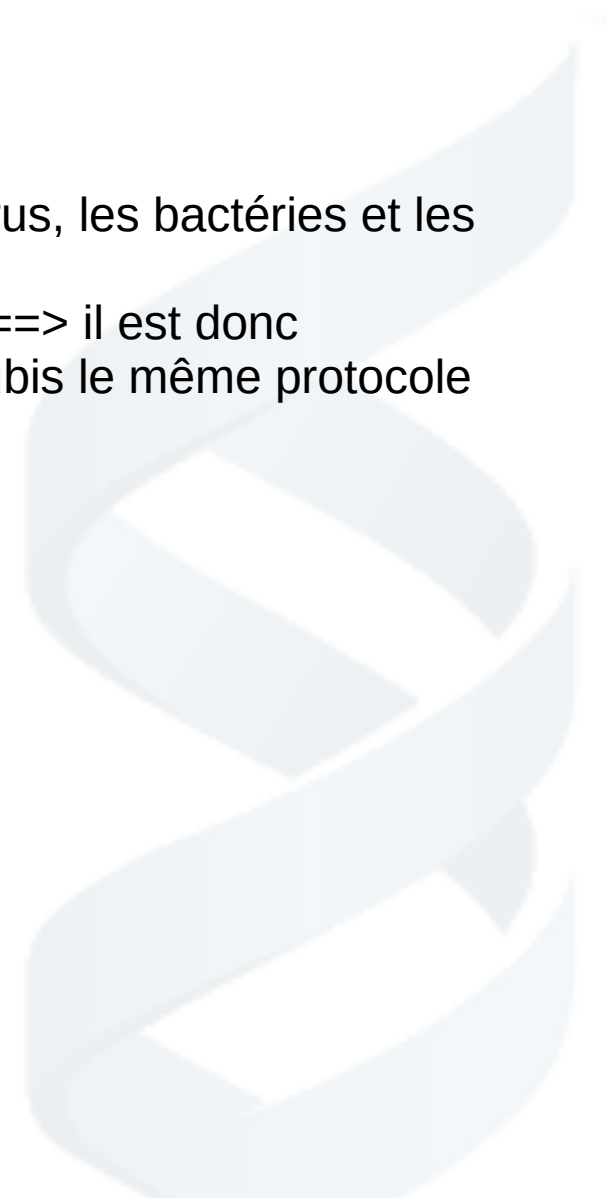


# 1. Introduction

## Méta-génomique (la partie biologique)

### Échantillonnage et extraction de l'ADN :

- Il dépend du milieu analysé (eau, sol, excréments...)
- Un filtre sur la taille est souvent fait pour discriminer les virus, les bactéries et les eucaryotes unicellulaires.
- Le stockage, l'exposition à l'oxygène entraînent des biais ==> il est donc indispensable que les échantillons à comparer aient tous subis le même protocole d'extraction.



# 1. Introduction

## Méta-génomique (la partie biologique)

### **Échantillonnage et extraction de l'ADN :**

- Il dépend du milieu analysé (eau, sol, excréments...)
- Un filtre sur la taille est souvent fait pour discriminer les virus, les bactéries et les eucaryotes unicellulaires.
- Le stockage, l'exposition à l'oxygène entraînent des biais ==> il est donc indispensable que les échantillons à comparer aient tous subis le même protocole d'extraction.

### **Préparation des librairies :**

- Selon la quantité d'ADN extraite, il est possible qu'une PCR soit nécessaire
- L'amplification est à éviter à cause des biais importants qu'elle engendre

# 1. Introduction

## Méta-génomique (la partie biologique)

### **Échantillonnage et extraction de l'ADN :**

- Il dépend du milieu analysé (eau, sol, excréments...)
- Un filtre sur la taille est souvent fait pour discriminer les virus, les bactéries et les eucaryotes unicellulaires.
- Le stockage, l'exposition à l'oxygène entraînent des biais ==> il est donc indispensable que les échantillons à comparer aient tous subis le même protocole d'extraction.

### **Préparation des librairies :**

- Selon la quantité d'ADN extraite, il est possible qu'une PCR soit nécessaire
- L'amplification est à éviter à cause des biais importants qu'elle engendre

### **Séquençage :**

- L'assemblage étant une étape complexe, des reads longs (ONT : Oxford Nanopore Technology – encore trop d'erreurs - , PACBIO HiFi ou linked reads ) sont à considérer
- Mais le coût est très important et les outils en cours de développement.

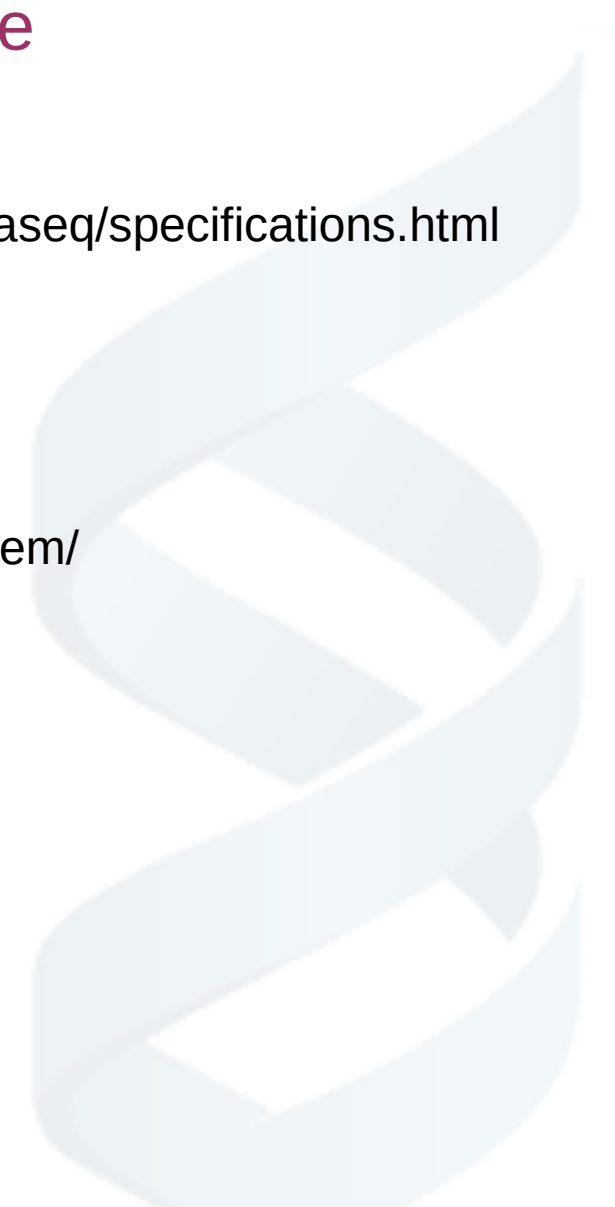
Actuellement un séquençage Illumina NovaSeq est le plus souvent pratiqué.

- Ce qui permet une grande profondeur de séquençage mais les reads sont courtes (2x150 pb, quelques milliards de paires de reads par flowcell, jusqu'à 10 pour le NovaSeq)
- Intéressant lorsqu'on veut avoir accès aux espèces rares dans le milieu étudié.

# 1. Introduction

## Les principales technologies de séquençage

- **Illumina** : NovaSeq  
<https://emea.illumina.com/systems/sequencing-platforms/novaseq/specifications.html>
- **Pacbio** HiFi vs CLS  
<https://www.pacb.com/technology/hifi-sequencing/>  
<https://www.pacb.com/technology/hifi-sequencing/sequel-system/>  
<https://www.pacb.com/revio/>
- **ONT** (Oxford nanopore technologie)  
<https://nanoporetech.com/how-it-works/basecalling>  
<https://nanoporetech.com/q20plus-chemistry>



1. Introduction
- 2. Applications**
3. Analyses méta-omiques
  - 3.1. La méta-génétique
  - 3.2. La méta-génomique
4. Analyses exploratoires
5. Impact carbone du calcul



## 2. Applications

### **Science environnementale :**

- Cycles des éléments fondamentaux (carbone, azote...)
- Contrôle de la pollution voir dépollution
- Ecologie / évolution



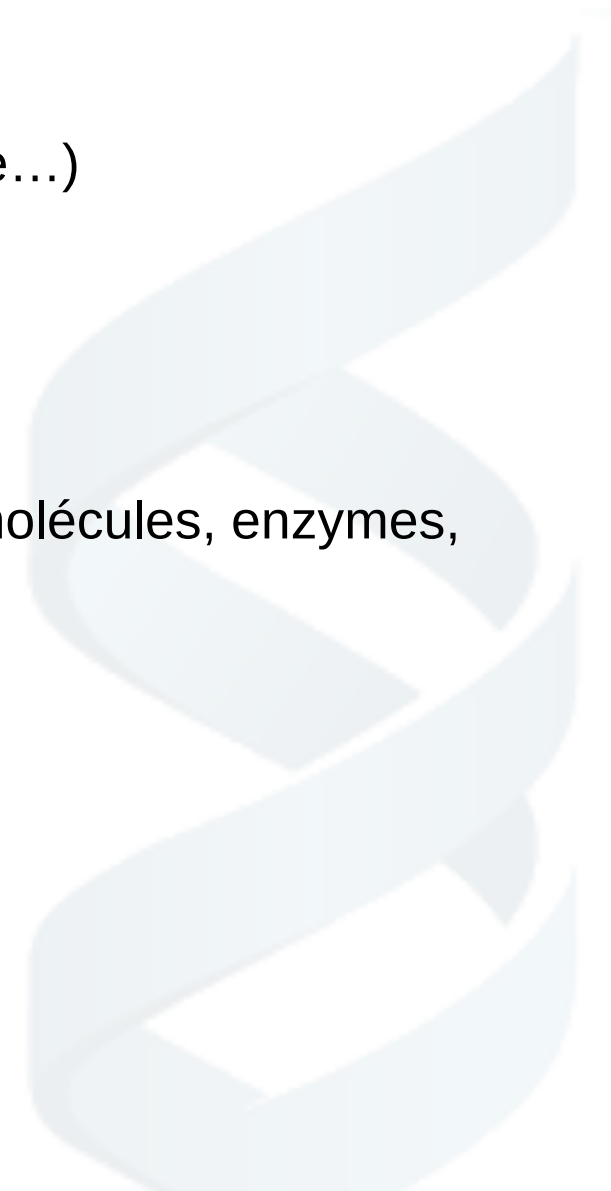
## 2. Applications

### Science environnementale :

- Cycles des éléments fondamentaux (carbone, azote...)
- Contrôle de la pollution voir dépollution
- Ecologie / évolution

### Applications industrielles :

- épuration des eaux usées
- prospection biologique (pour trouver de nouvelles molécules, enzymes, antibiotiques etc..)
- nouvelles biosynthèses
- fermentations (yaourts...)



## 2. Applications

### **Science environnementale :**

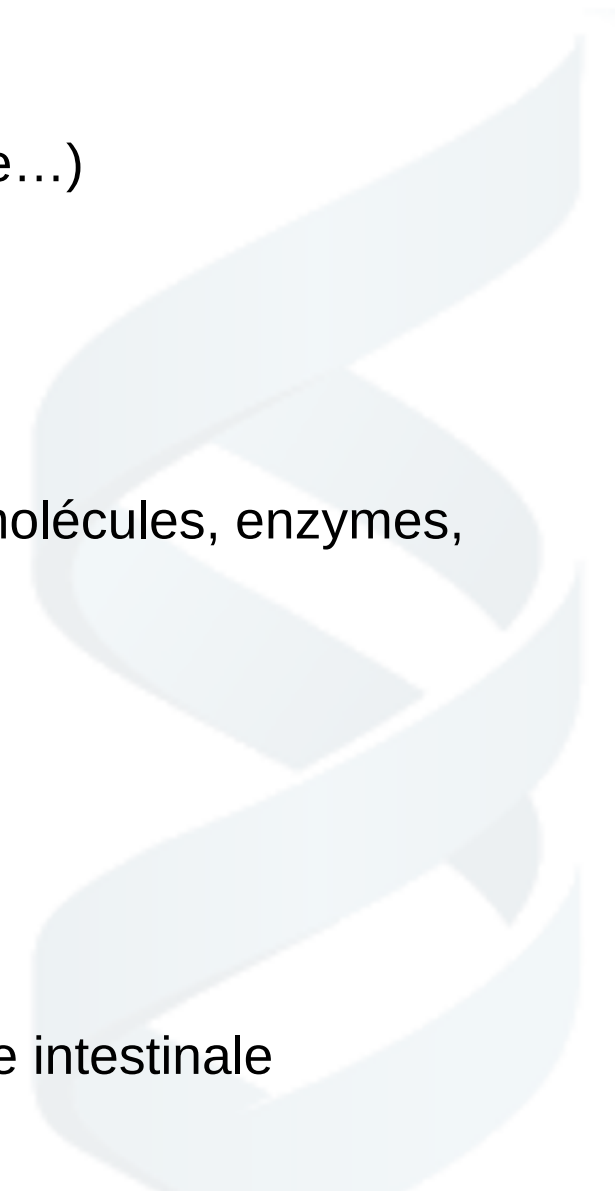
- Cycles des éléments fondamentaux (carbone, azote...)
- Contrôle de la pollution voir dépollution
- Ecologie / évolution

### **Applications industrielles :**

- épuration des eaux usées
- prospection biologique (pour trouver de nouvelles molécules, enzymes, antibiotiques etc..)
- nouvelles biosynthèses
- fermentations (yaourts...)

### **Santé humaine et animale :**

- protection contre les pathogènes
- cancer
- absorption des nutriments
- certaines maladies chroniques (parodontite, maladie intestinale inflammatoire, par ex. La maladie de Crohn)

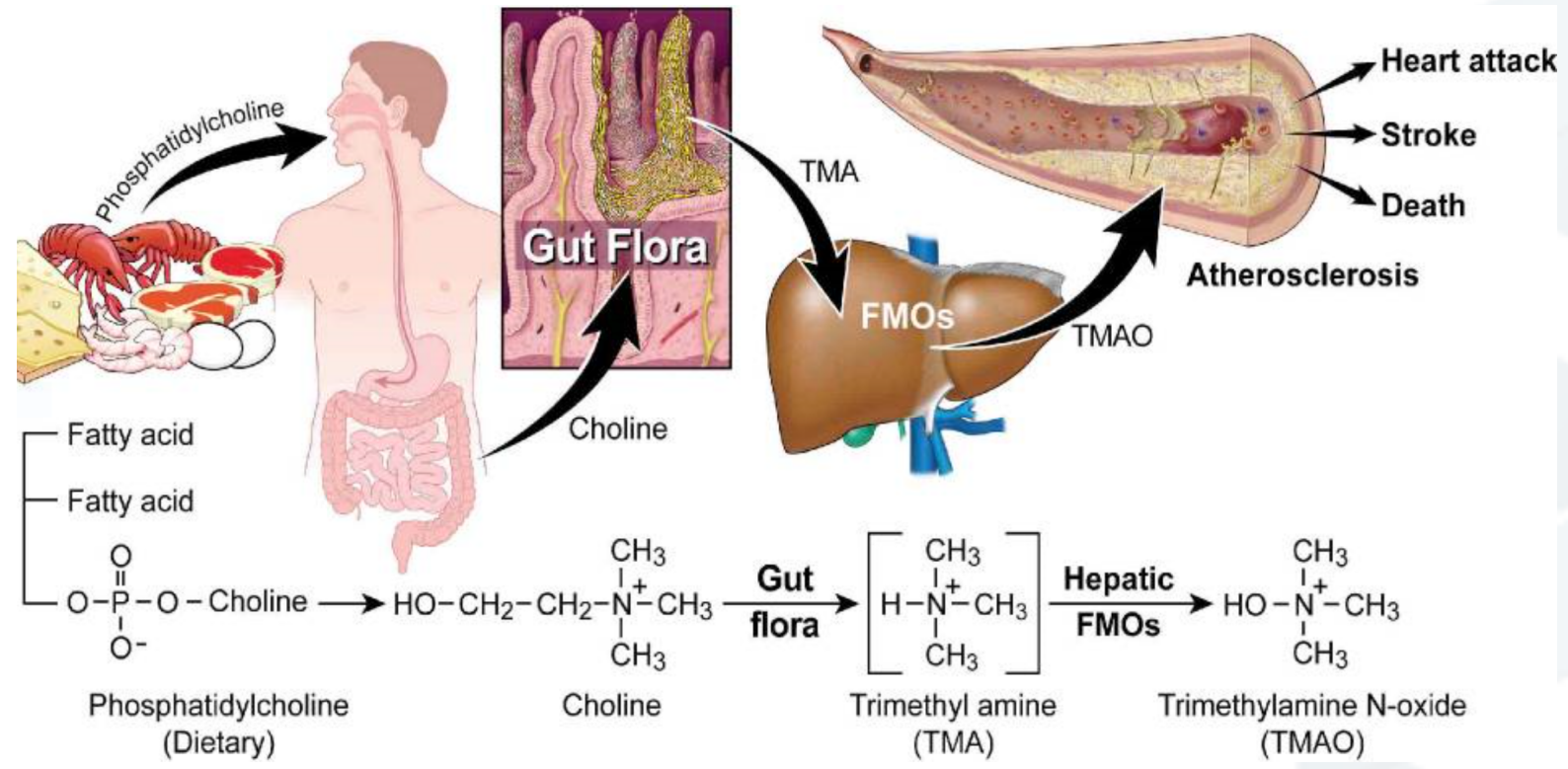




# 2. Applications

## Quelques exemples

Impact sur les maladies cardiovasculaires  
(Wang et al., 2011)



## 2. Applications

### Quelques exemples

Influence sur le cancer colorectal  
(Sears, C. L., & Garrett, W. S., 2014).

**Table 1. Criteria for Disease Causation: Human Colorectal Cancer and Putative Bacterial Protagonists**

Criteria <sup>a</sup>	<i>S. gallolyticus</i>	ETBF	<i>E. faecalis</i>	<i>E. coli</i>	<i>F. nucleatum</i>
Epidemiology <sup>b</sup>	+	+	–	+	+
Measurable immunological responses <sup>c</sup>	+	–	–	–	–
Experimental disease reproduction <sup>d</sup>	–	+	+	+ <sup>e</sup>	+
Biological plausibility <sup>f</sup>	±	+	±	+	+
Elimination or modification of agent prevents disease <sup>g</sup>	–	–	–	–	–

Presence or absence of data is noted by + (present) or – (absent); ± denotes overall data are variable.

<sup>a</sup>Adapted from [Evans \(1976\)](#) and [Fredericks and Relman \(1996\)](#).

<sup>b</sup>Epidemiology encompasses several types of evidence, including prevalence, exposure, or incidence of disease significantly higher in those exposed to the putative cause than controls; data comparing cases and controls should show consistency and strength of association; a range of controls should be evaluated to assess specificity of the epidemiologic association; temporality (exposure antedates disease development).

<sup>c</sup>Only data assessing human immunologic responses are considered.

<sup>d</sup>Experimental disease induction refers to animal models demonstrating increased colon carcinogenesis by the listed bacterium.

<sup>e</sup>Experimental model data are only available for *E. coli* possessing the pks island.

<sup>f</sup>Biologic plausibility reflects the authors' judgment of the strength of the data available at present regarding the potential role of the bacterial protagonist in human CRCs.

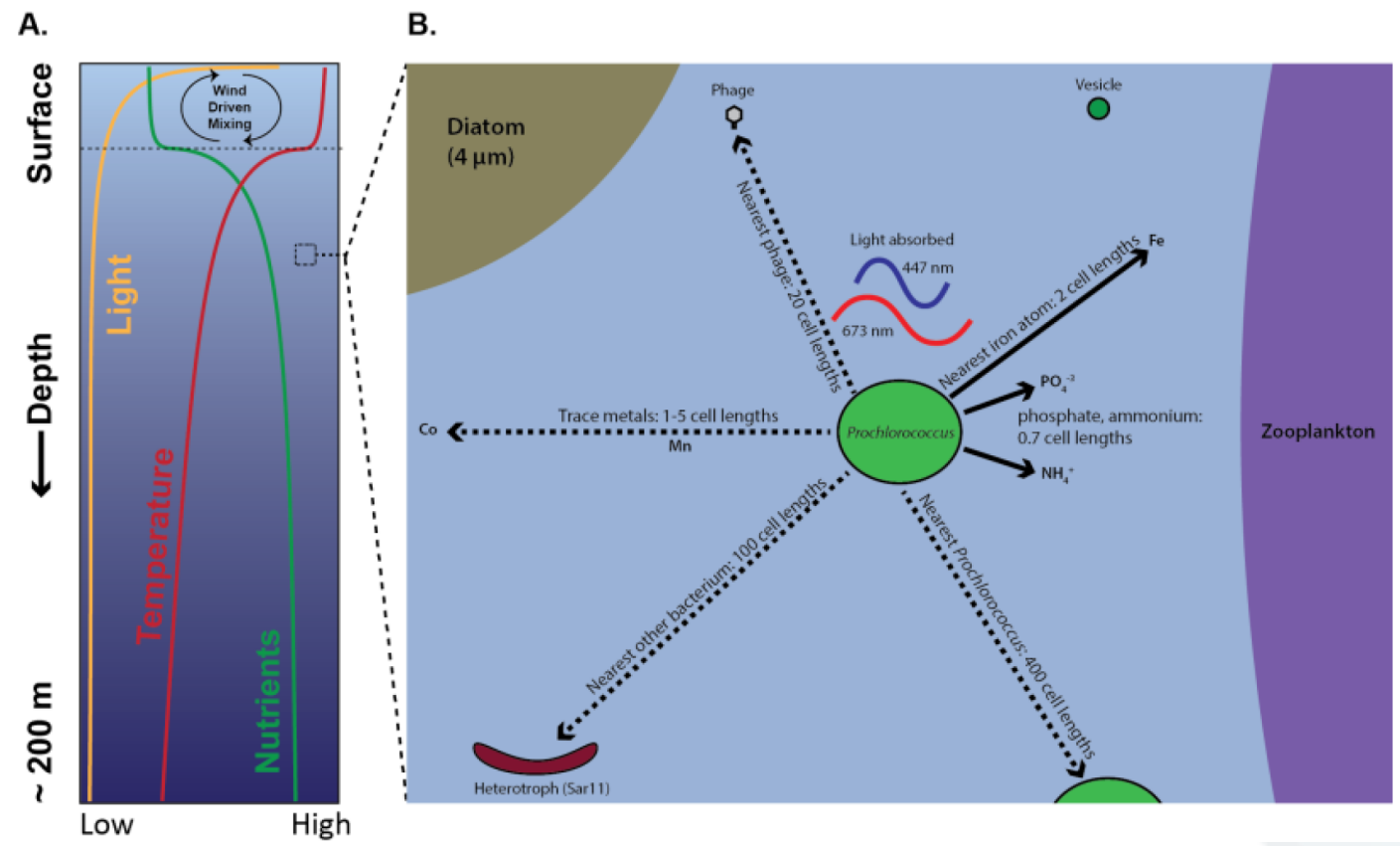
<sup>g</sup>Elimination or modification of agent prevents disease refers to human studies such as use of antibiotics, probiotics, or vaccines to prevent disease. As yet, no such studies are reported for these bacteria and CRC.

# 2. Applications

## Quelques exemples

Environnement : *Prochlorococcus*  
 (Biller et al, 2014).

L'organisme photosynthétique le plus abondant et le plus petit sur terre.



## 2. Applications

### Quelques exemples

Environnement : *Prochlorococcus*  
(Biller et al, 2014).

ITS : Internal transcribed spacer

	clade II/III <sup>37</sup>		
LLIV	eMIT9313 <sup>29</sup> , High-B/A <i>Prochlorococcus</i> clade IV <sup>37</sup>	MIT9303, MIT9313, MIT0701	Typically most abundant near the base of the euphotic zone; highly susceptible to light shock <sup>35,37</sup> .
LLV		None	Found maximally abundant in the lower euphotic zone of oxygen minimum zones when oxygen depleted layers extend into the upper water column <sup>56</sup> .
LLVI		None	Found maximally abundant in the lower euphotic zone of oxygen minimum zones when oxygen depleted layers extend into the upper water column <sup>56</sup> .
LLVII	NCI <sup>36</sup>	None	Sequences were identified in the lower euphotic zone of subtropical waters; little is known about this clade <sup>36</sup> .

<sup>a</sup> Originally defined as separate clades<sup>37</sup>, the LLII and LLIII clades are now grouped because their separation is not well resolved phylogenetically.

<sup>b</sup> Two publications<sup>49,50</sup> assigned the names HNLC1 and HNLC2 to different clades; moving forward, we suggest the use of the HLIII and HLIV nomenclature to refer to these clades<sup>47,48</sup>.

<sup>c</sup> For more information on these and other strains, see<sup>10,44,60</sup> and references therein.

Table 1. The major clades of *Prochlorococcus* as defined by rRNA ITS sequences.

Clade Name	Alternate names in the literature for the same ribotype	Representative cultured strains <sup>c</sup>	Habitat where relatively more abundant, and/or where isolated
HLI	eMED4 <sup>29</sup> , Low-B/A <i>Prochlorococcus</i> clade I <sup>37</sup>	MED4, MIT9515	Isolated from the upper/mid euphotic zone, typically from the subtropical ocean. Their distribution is shifted to higher latitudes, due to a relatively low temperature growth optimum <sup>31,33-35</sup> .
HLII	eMIT9312 <sup>29</sup> , Low-B/A <i>Prochlorococcus</i> clade II <sup>37</sup>	AS9601, MIT9215, MIT9312, SB	Often found throughout the euphotic zone; typically among the most abundant <i>Prochlorococcus</i> group in the water column. Especially abundant at lower latitudes, due to a relatively high temperature growth optimum <sup>31,33-35</sup> .
HLIII	HNLC1 <sup>47,50</sup> , HNLC2 <sup>49,b</sup>	None	Sequences from high nutrient, but low chlorophyll containing equatorial waters, typically between 10 °N – 10 °S in the Pacific and Indian oceans. These regions are typically limited by iron availability, and sequence data suggests that these cells have adaptations for reducing cellular iron requirements <sup>47-50</sup> .
HLIV	HNLC1 <sup>49</sup> , HNLC2 <sup>47,50,b</sup>	None	Sequences from high nutrient, but low chlorophyll containing equatorial waters, typically between 10 °N – 10 °S in the Pacific and Indian oceans. These regions are typically limited by iron availability, and sequence data suggests that these cells have adaptations for reducing cellular iron requirements <sup>47-50</sup> .
HLV		None	HLV sequences have been found in surface equatorial waters typically limited by iron availability. Physiological distinctions between the HLIII, HLIV and HLV clades are not known <sup>47</sup> .
HLVI		None	Sequences were identified in the mid/lower euphotic zone (75-150m) of the South China Sea. This group has been postulated to have an intermediate light optimum <sup>47</sup> .
LLI	eNATL2A <sup>29</sup> , High-B/A <i>Prochlorococcus</i> clade I <sup>37</sup>	NATL1A, NATL2A, PAC1	Typically most abundant in the middle euphotic zone of stratified waters. Unlike other LL clades, they often remain abundant in mixed waters throughout the water column due to their ability to tolerate light shock <sup>35,37</sup> .
LLII/III <sup>a</sup>	eSS120/eMIT921 <sup>29</sup> , High-B/A <i>Prochlorococcus</i>	MIT9211, SS120	Typically found in the middle/lower euphotic zone <sup>35,37,44</sup> .

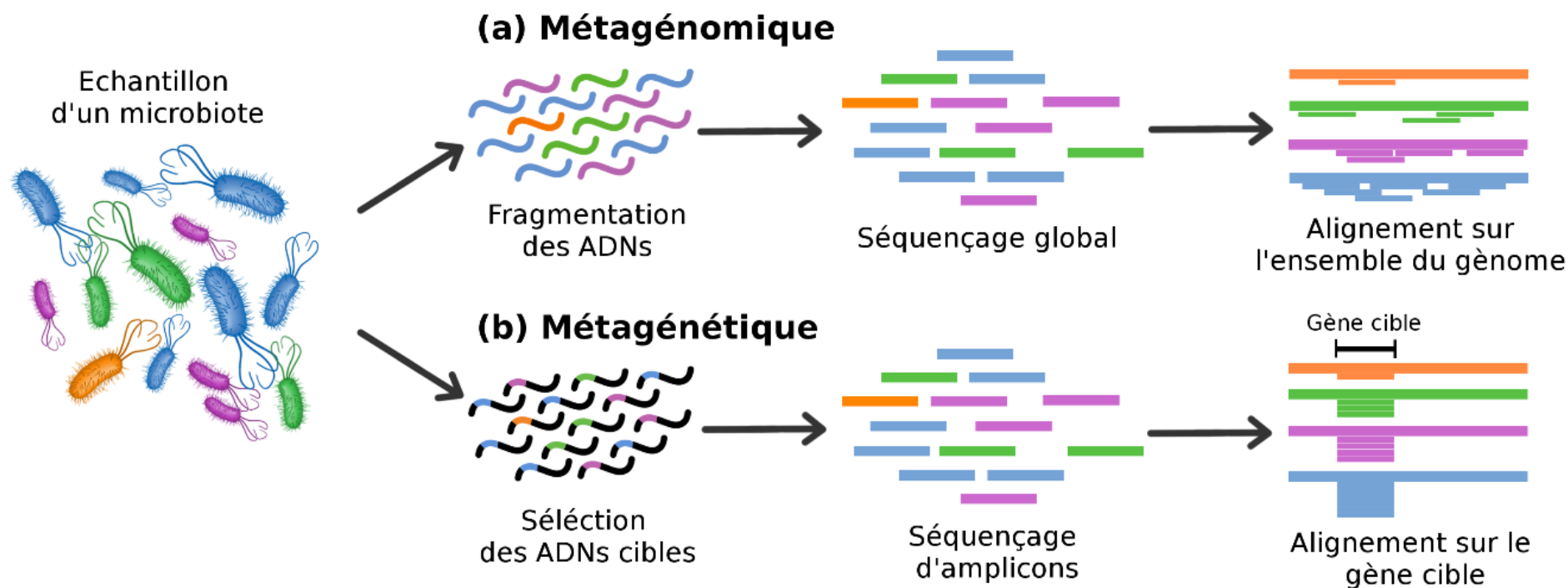
1. Introduction
2. Applications
- 3. Analyses méta-omiques**
  - 3.1. La méta-génétique**
  - 3.2. La méta-génomique**
4. Analyses exploratoires
5. Impact carbone du calcul





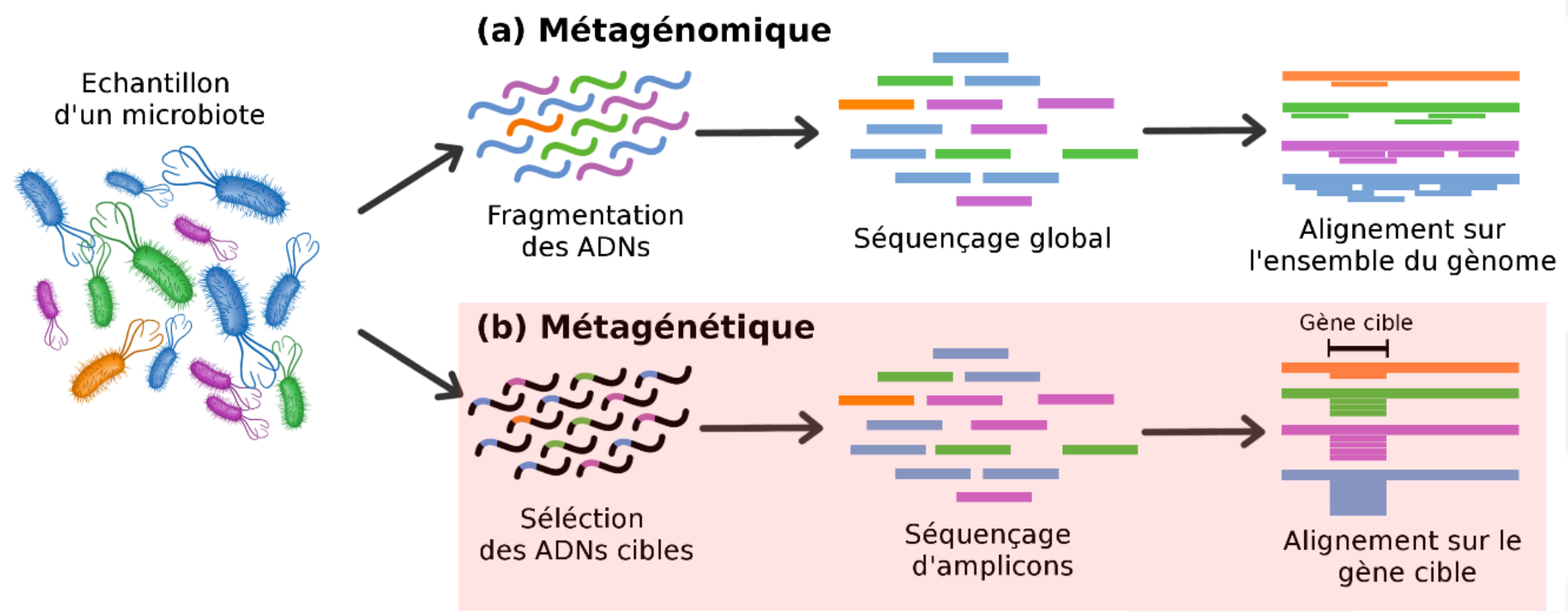
# 3. Analyses méta-omiques

## Méta-génétique & méta-génomique



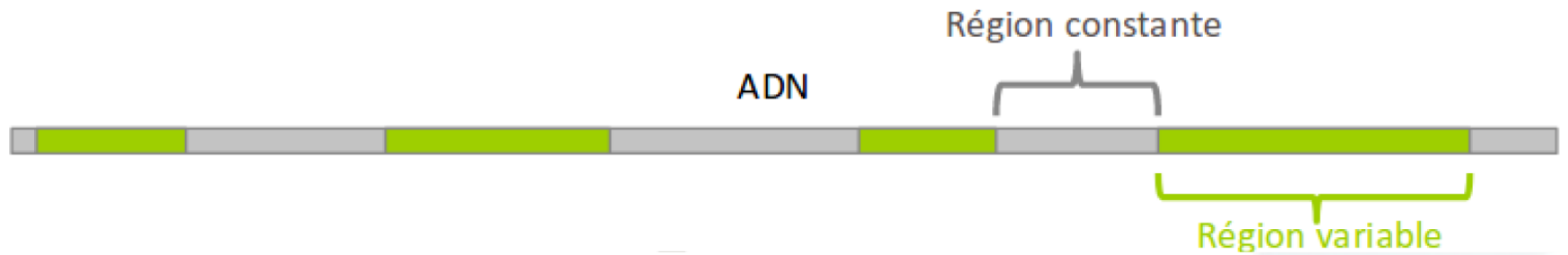
# 3. Analyses méta-omiques

## Méta-génétique & méta-génomique



# 3. Analyses méta-omiques

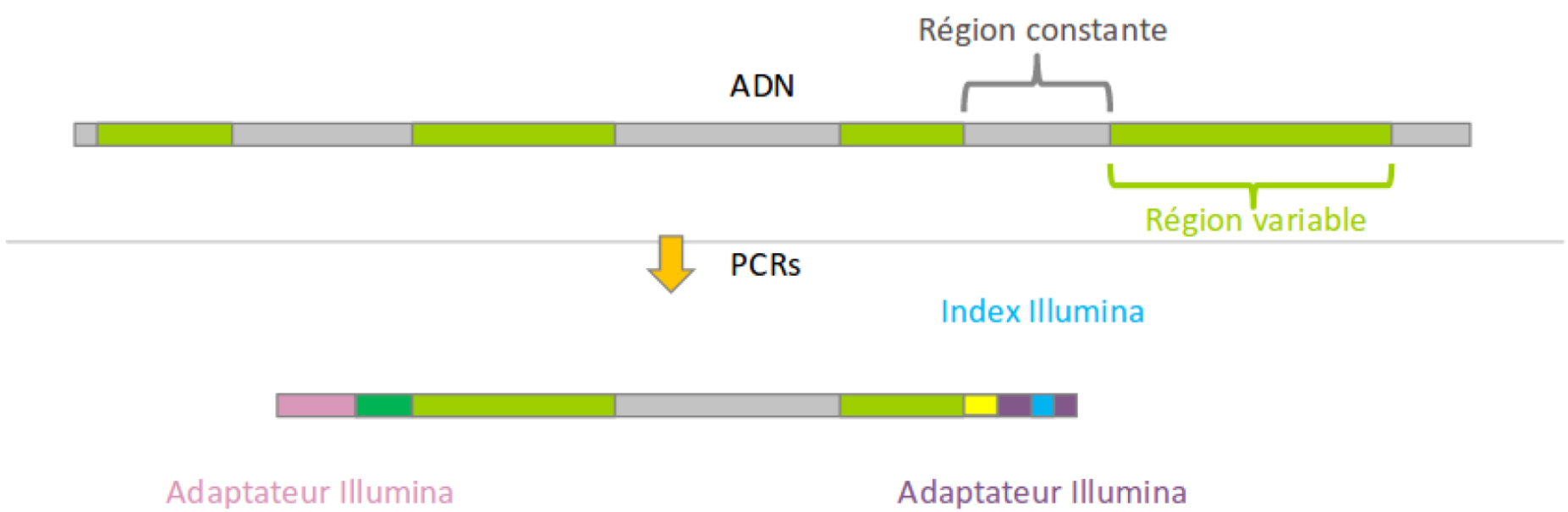
La méta-génétique – les données





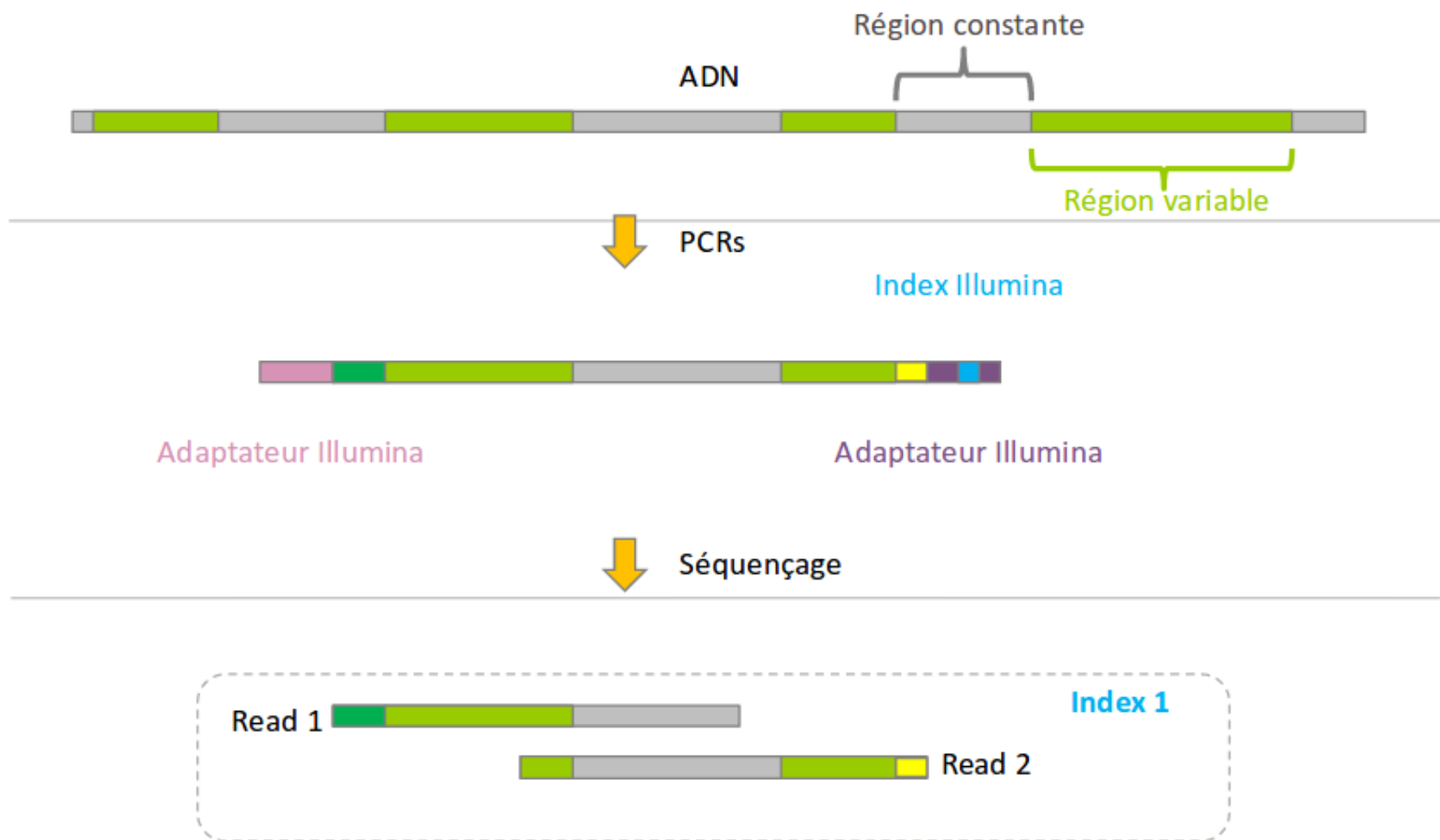
# 3. Analyses méta-omiques

## La méta-génétique – les données



# 3. Analyses méta-omiques

## La méta-génétique – les données



# 3. Analyses méta-omiques

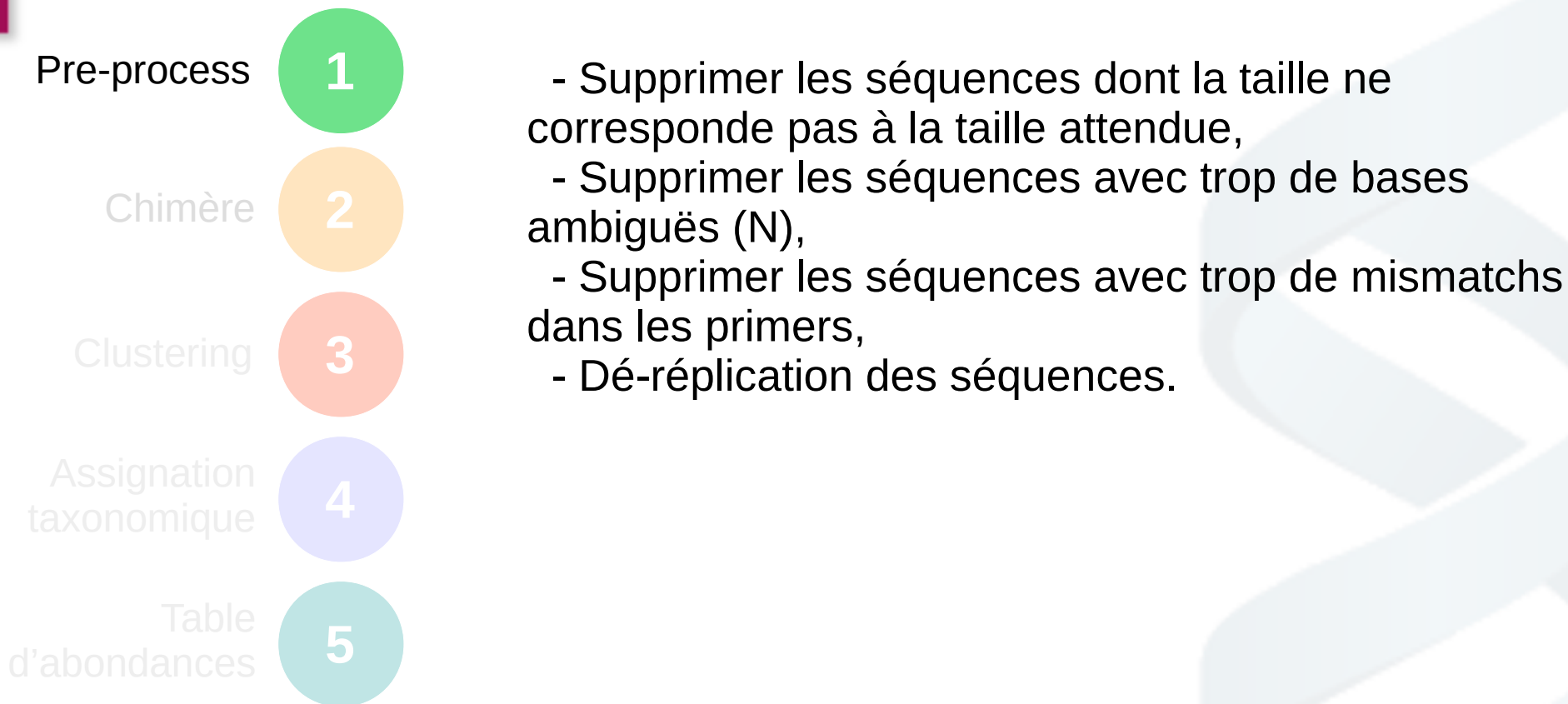
## La méta-génomique – le workflow

- Pre-process **1**
- Chimère **2**
- Clustering **3**
- Assignment taxonomique **4**
- Table d'abondances **5**



# 3. Analyses méta-omiques

## La méta-génomique – le workflow



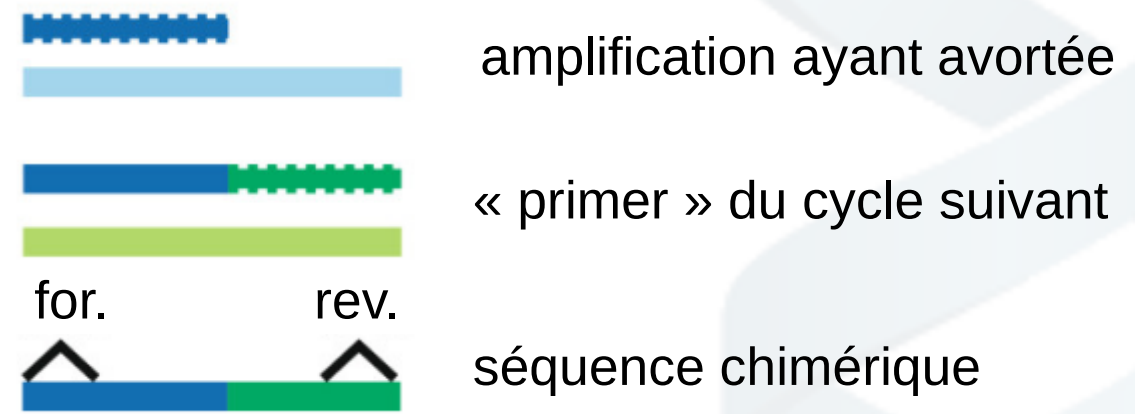
# 3. Analyses méta-omiques

## La méta-génétique – le workflow

- Pre-process 1
- Chimère 2
- Clustering 3
- Assignment taxonomique 4
- Table d'abondances 5

**Définition :** principalement produits lors de la phase de PCR lorsqu'un amplicon avorté joue le rôle de primeur pour un template hétérologue. La séquence a une taille similaire et contient le primer forward et reverse.

**Schloss, 2011 :** 5 à 45 % de lectures



# 3. Analyses méta-omiques

## La méta-génomique – le workflow

### Détection :

- recherche des séquences parentes
- validation croisée entre échantillon

Pre-process

1

Chimère

2

Clustering

3

Assignment  
taxonomique

4

Table  
d'abondances

5



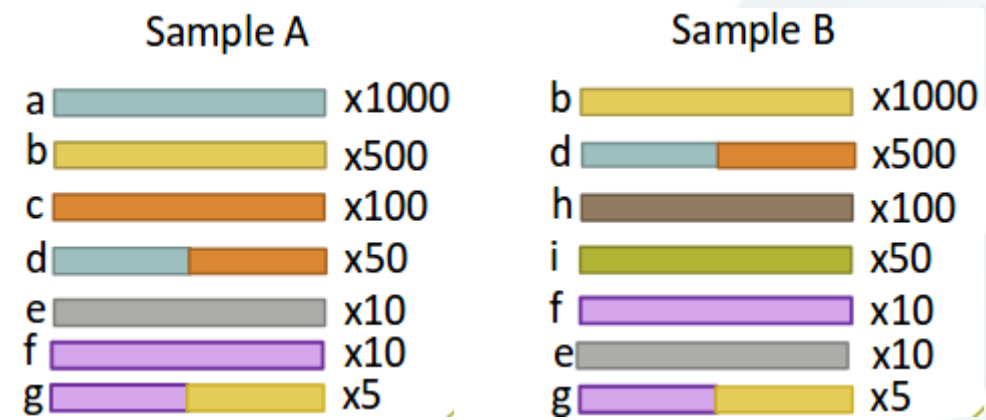
# 3. Analyses méta-omiques

## La méta-génétique – le workflow



### Détection :

- recherche des séquences parentes
- validation croisée entre échantillon



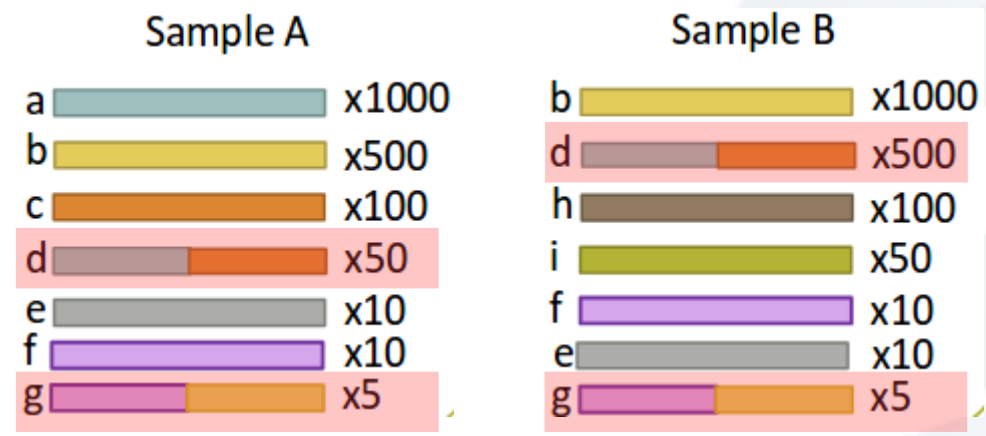
# 3. Analyses méta-omiques

## La méta-génétique – le workflow



### Détection :

- recherche des séquences parentes
- validation croisée entre échantillon





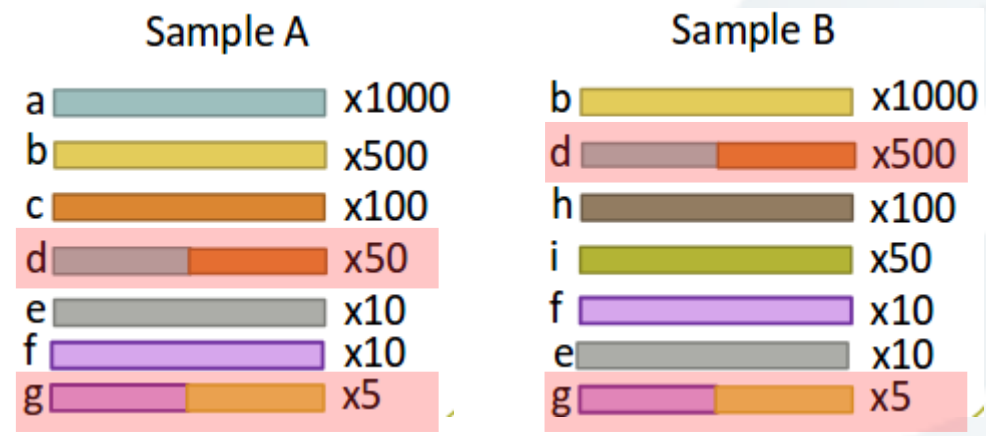
# 3. Analyses méta-omiques

## La méta-génétique – le workflow



### Détection :

- recherche des séquences parentes
- validation croisée entre échantillon



**d** n'est pas une chimère  
**g** est une chimère

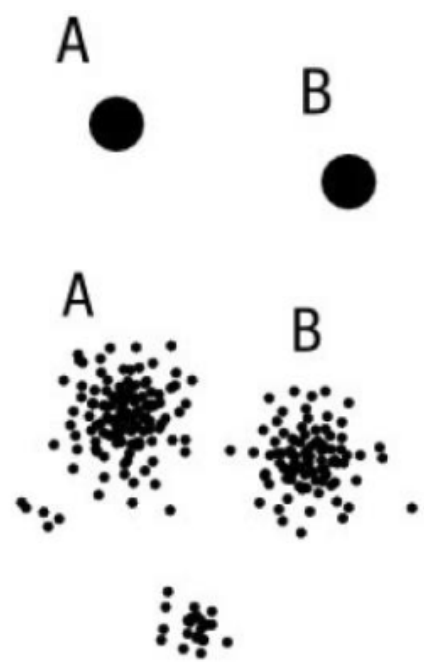
**Outils :** Uchime, ChimeraSlayer, Uparse, CATCh

# 3. Analyses méta-omiques

## La méta-génétique – le workflow

- Pre-process 1
- Chimère 2
- Clustering** 3
- Assignation taxonomique 4
- Table d'abondances 5

**Objectif :** regrouper les séquences appartenant à la même Unité Taxonomique Opérationnelle (OTU) ~ espèce.



Ce qui est attendu

La réalité

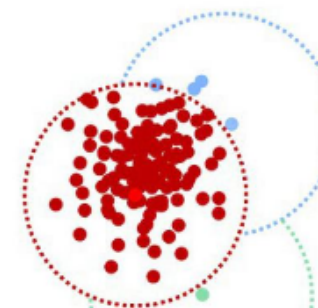
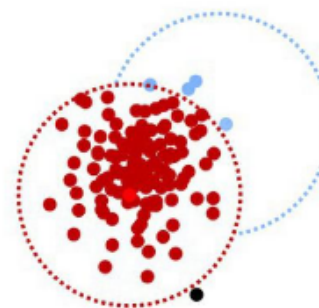
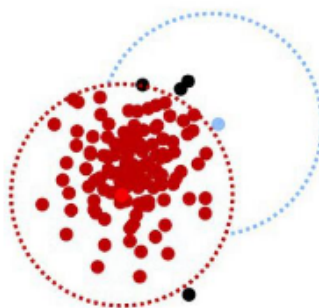
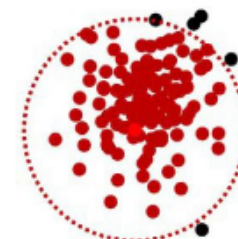
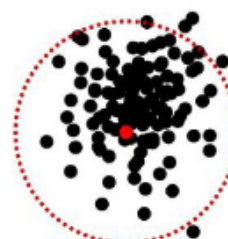
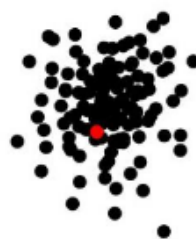
**Causes :** Variabilité naturelle, erreurs techniques, contaminations, chimères, ...

# 3. Analyses méta-omiques

## La méta-génétique – le workflow

### Algorithmes gloutons

- Pre-process **1**
- Chimère **2**
- Clustering **3**
- Assignation taxonomique **4**
- Table d'abondances **5**

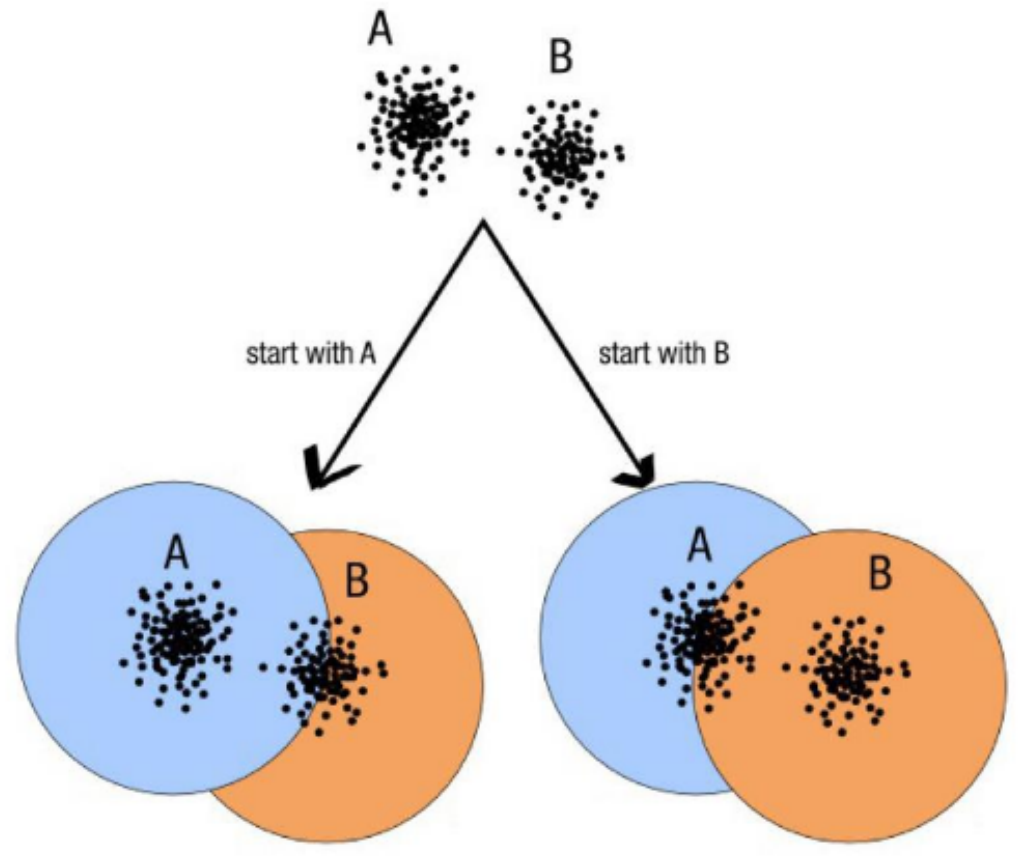


# 3. Analyses méta-omiques

## La méta-génétique – le workflow

**Algorithmes gloutons** : dépendant de l'initialisation

- Pre-process 1
- Chimère 2
- Clustering** 3
- Assignation taxonomique 4
- Table d'abondances 5



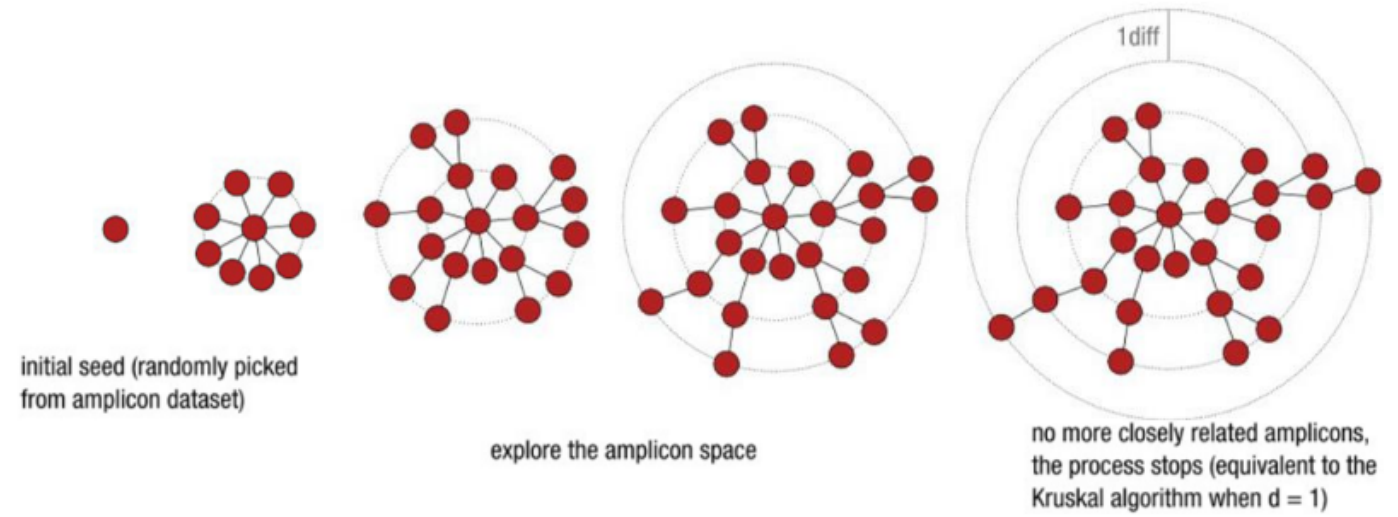
# 3. Analyses méta-omiques

## La méta-génétique – le workflow

**SWARM** (Mahé, 2014) : clustering robuste et rapide adapté à de grand jeu de données.

- Pre-process 1
- Chimère 2
- Clustering 3
- Assignment taxonomique 4
- Table d'abondances 5

	ACGT	ACGT	ACGT
	AGGT	A - GT	A - - T
differences	1	1	2



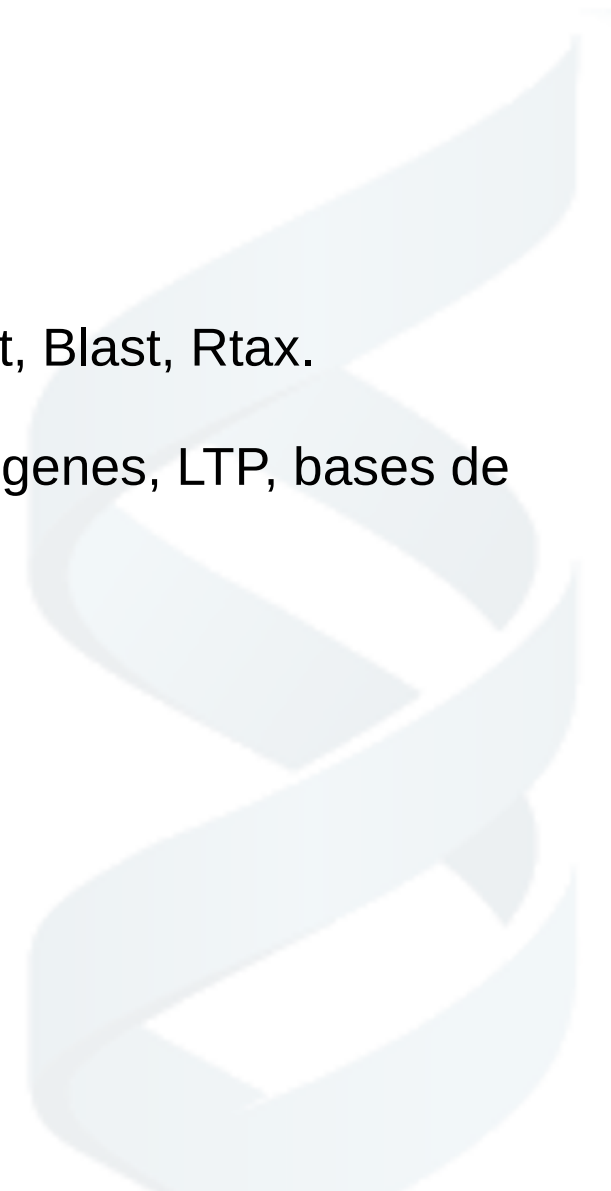
# 3. Analyses méta-omiques

## La méta-génomique – le workflow



**Outils** : RDPclassifier, Megablast, Blast, Rtax.

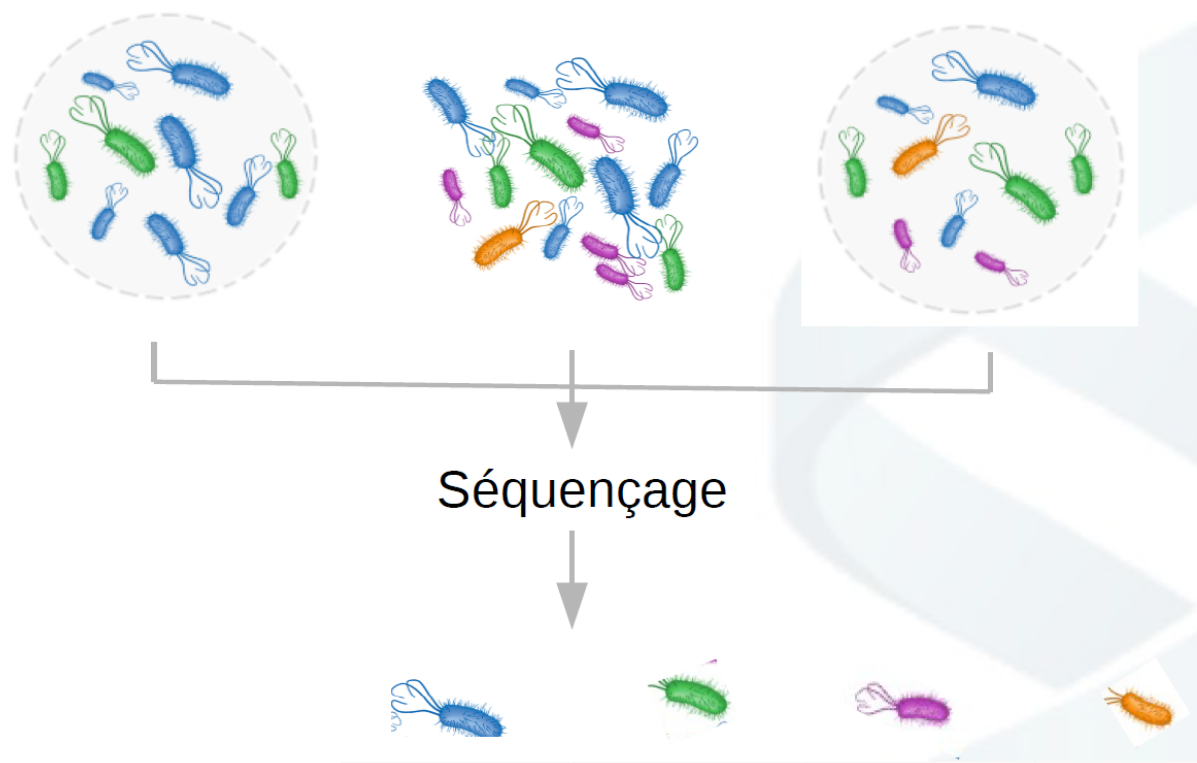
**Base de données** : Silva, Greengenes, LTP, bases de données dédiées.



# 3. Analyses méta-omiques

## La méta-génétique – le workflow

- Pre-process **1**
- Chimère **2**
- Clustering **3**
- Assignation taxonomique **4**
- Table d'abondances **5**



échantillon 1	6	3	0	0
échantillon 2	6	3	4	1
échantillon 3	3	3	2	1

# 3. Analyses méta-omiques

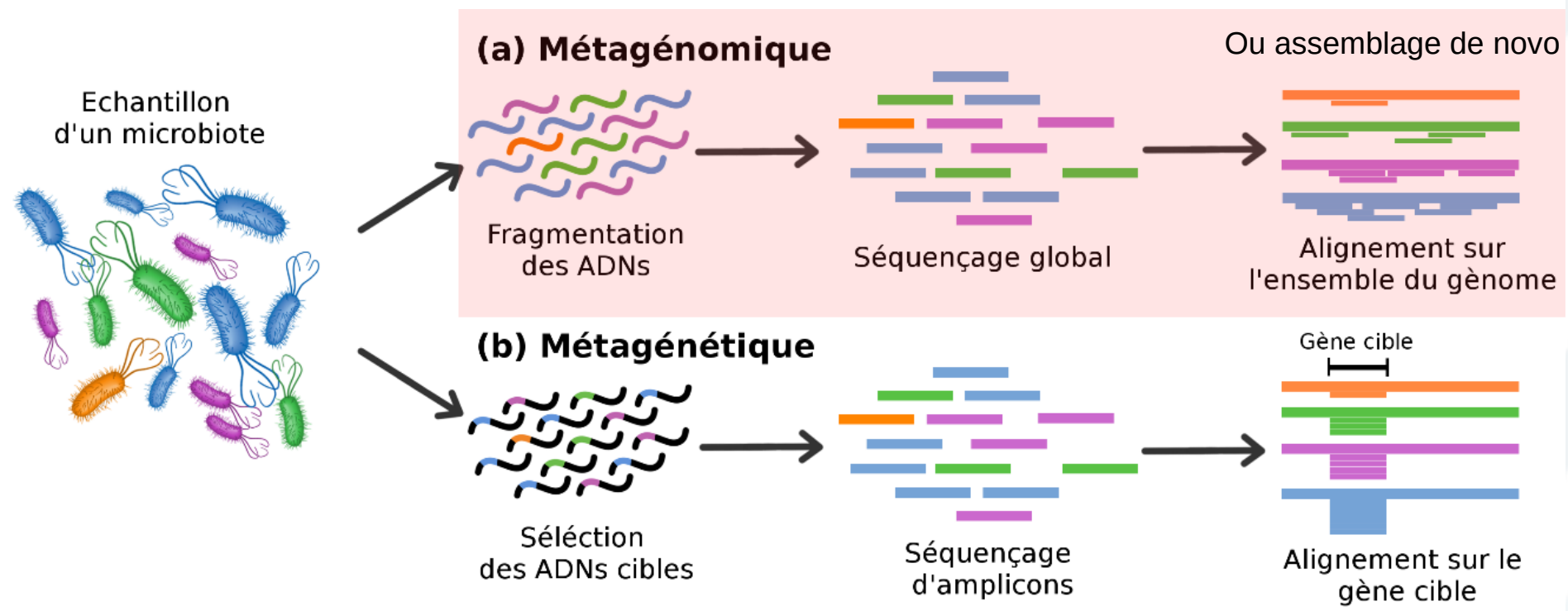
## La méta-génétique – les outils

QIIME	
UPARSE	<a href="https://www.drive5.com/uparse/">https://www.drive5.com/uparse/</a>
MOTHUR	<a href="https://www.mothur.org">https://www.mothur.org</a>
MG-RAST	<a href="http://metagenomics.anl.gov">http://metagenomics.anl.gov</a>
FROGS	<a href="http://frogs.toulouse.inra.fr">http://frogs.toulouse.inra.fr</a>



# 3. Analyses méta-omiques

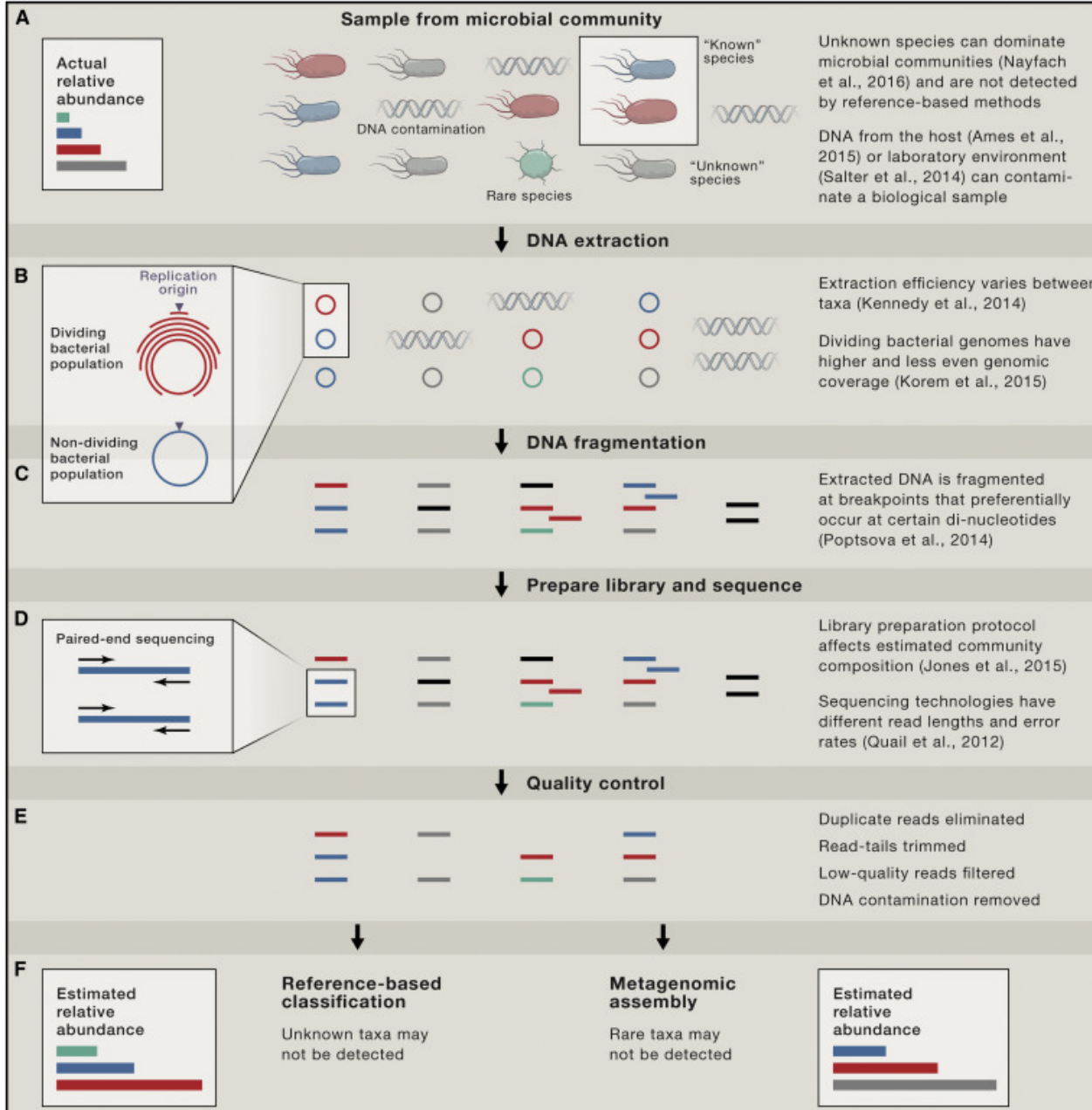
## Méta-génétique & méta-génomique



<https://www.youtube.com/watch?v=DIQTXdb2rhg&list=PLOPiWVjg6aTzsA53N19YqJQeZpSCH9QPc&index=23>  
 Dan Knights Videos

# 3. Analyses méta-omiques

## La méta-génomique – les données



Nayfach, S., & Pollard, K. S. (2016).

# 3. Analyses méta-omiques

## La méta-génomique – le(s) workflow(s)

Le workflow d'analyse est beaucoup moins bien établi qu'en métagénétique, bien qu'il y en ait un florilège depuis quelques temps.

Il dépend de la question posée, de la quantité de données...

Les données brutes issues des analyses de métagénomiques sont complexes et bruitées.

Imaginez de l'ordre de ~1000 puzzles de millions de pièces avec des erreurs à reconstruire. De plus pour la plupart on ne connaît pas le modèle !

Extraire de l'information biologique de jeux de données de cette taille est un challenge et nécessite de grosses ressources de calcul.



# 3. Analyses méta-omiques

La méta-génomique – quelques étapes svt rencontrées

- 1 Pre-process
- 2 Assemblage  
Alignements
- 3 Annotation  
des contigs  
Comptages
- 4 Création du  
catalogue de  
gènes
- 5 Binning des  
contigs
- 6 Tables  
d'abondance  
taxonomiques et  
fonctionnelles



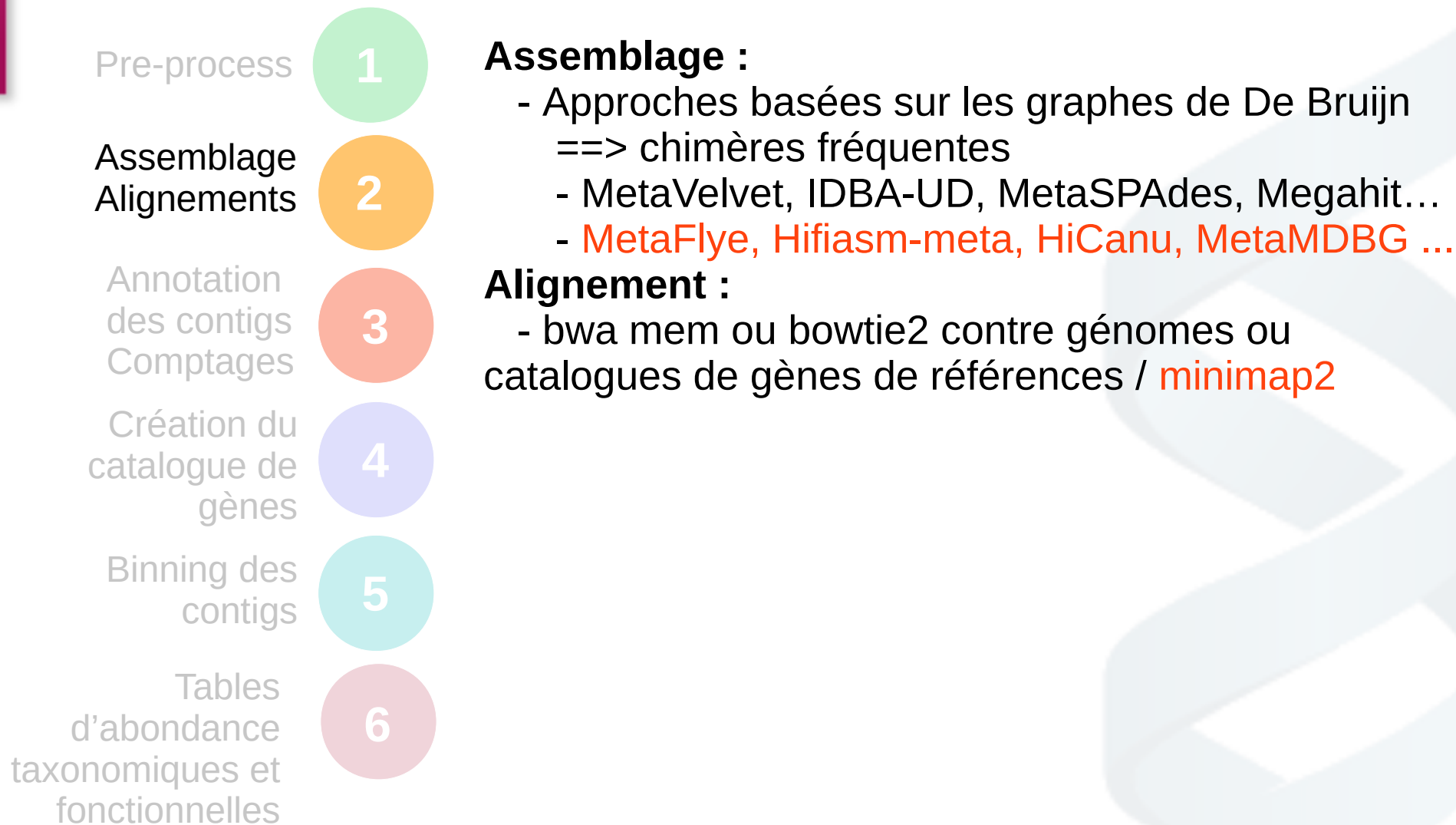
# 3. Analyses méta-omiques

La méta-génomique – quelques étapes svt rencontrées

Pre-process	1	<b>Short / long (HiFi) si outil spécifique</b> <b>Vérification de la qualité :</b> - fastQC
Assemblage Alignements	2	<b>Enlever les adaptateurs :</b> - cutadapt (SR)
Annotation des contigs Comptages	3	<b>Trimmer les bases sur leur qualité de séquençage :</b> - sickle / <b>Smrtlink (Pacbio, lors du séquençage)</b> <b>Vérifier rapidement la composition taxonomique à partir des reads :</b>
Création du catalogue de gènes	4	- kraken2, metaPhlan4, Kaiju... / <b>Megan-LR</b> (Huson et al. 2018), <b>Pb-metagenomics-tools</b>
Binning des contigs	5	
Tables d'abondance taxonomiques et fonctionnelles	6	

# 3. Analyses méta-omiques

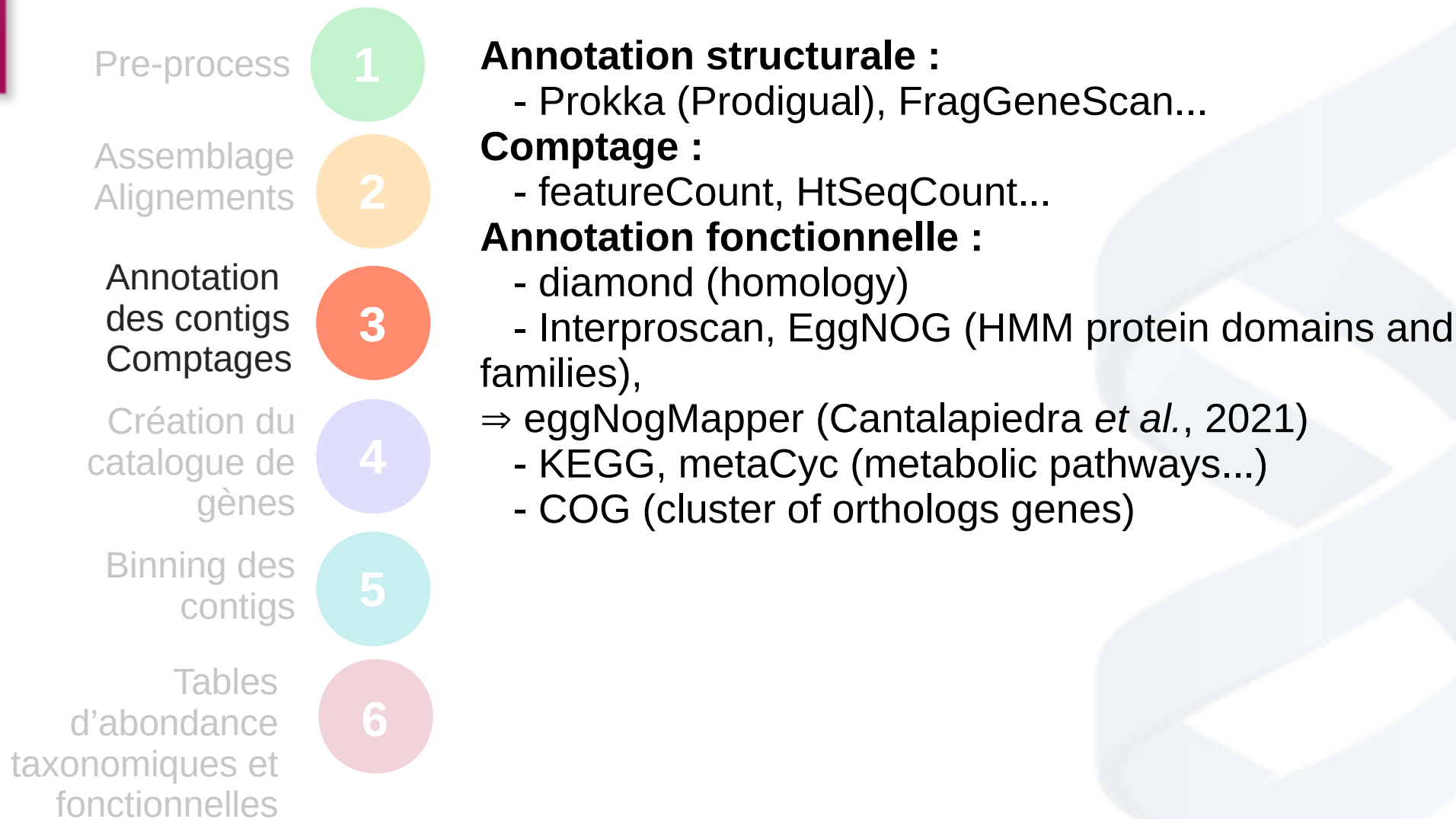
La méta-génomique – quelques étapes svt rencontrées





# 3. Analyses méta-omiques

## La méta-génomique – quelques étapes svt rencontrées



# 3. Analyses méta-omiques

La méta-génomique – quelques étapes svt rencontrées

- Pre-process **1**
- Assemblage  
Alignements **2**
- Annotation  
des contigs  
Comptages **3**
- Création du  
catalogue de  
gènes **4**
- Binning des  
contigs **5**
- Tables  
d'abondance  
taxonomiques et  
fonctionnelles **6**

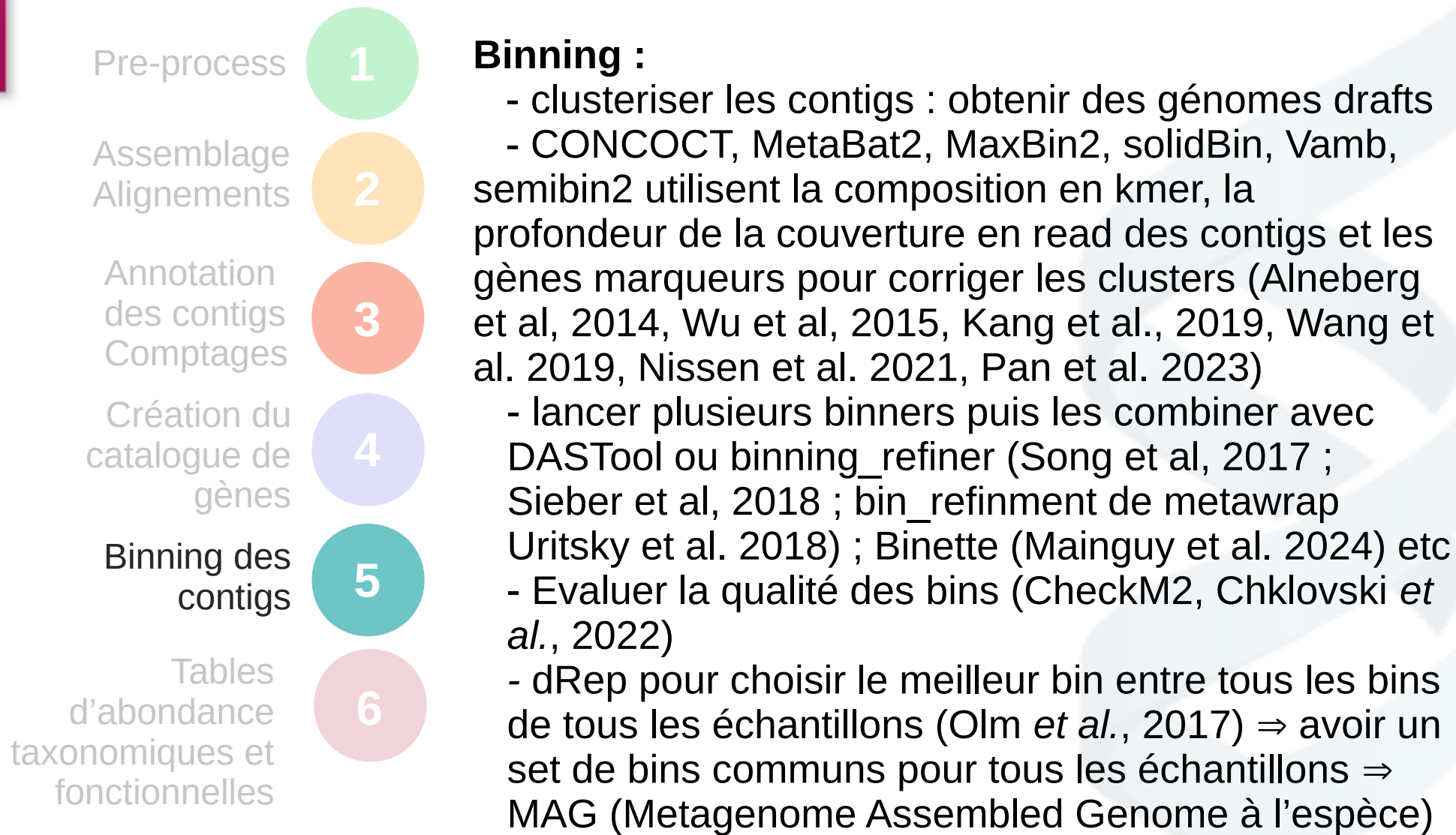
**Clustering des protéines :**  
- CD-Hit





# 3. Analyses méta-omiques

## La méta-génomique – quelques étapes svt rencontrées



# 3. Analyses méta-omiques

## La méta-génomique – quelques étapes svt rencontrées

- Pre-process

**1**

**Affiliation taxonomique (gènes puis contigs) :**

  - diamond : homology suivi d'un script algo LCA (lowest common ancestor) ex CAT et BAT (Bastiaan von Meijenfeldt *et al.* 2019)
- Assemblage  
Alignements

**2**

**Affiliation taxonomique (bins) :**

  - Gtdb-tk (Chaumeil *et al.*, 2022)
- Annotation  
des contigs  
Comptages

**3**

**Abondance fonctionnelle :**

A partir de la matrice des comptages des gènes :

  - profils d'abondance des orthologues (Kegg orthologie, COG, NOG)
  - profils d'abondance des pathways (Kegg or MetaCyc pathways or GO terms)
- Création du  
catalogue de  
gènes

**4**

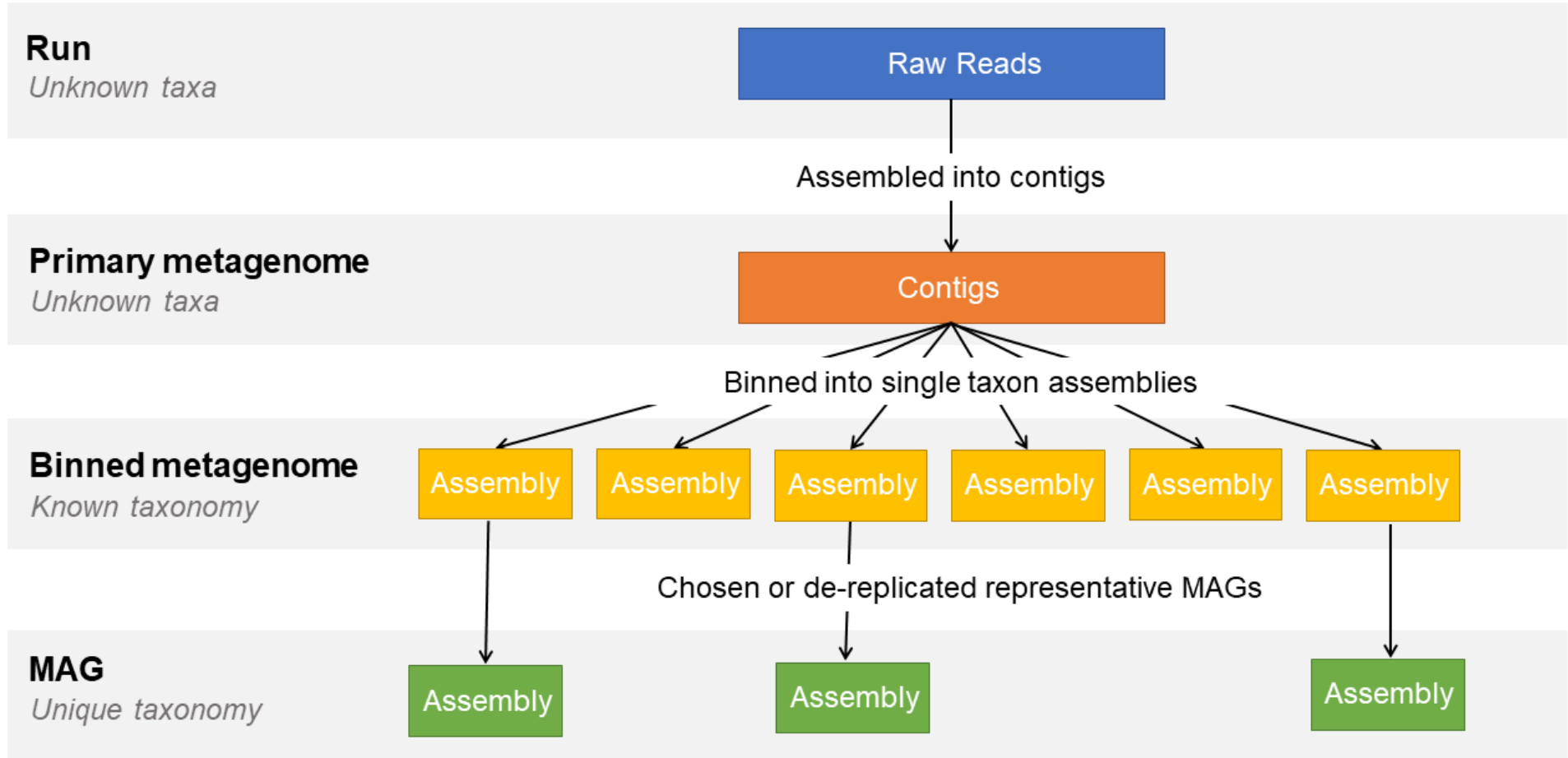
**Il s'agit en général d'abondance relative dans  
chaque échantillon**
- Binning des  
contigs

**5**
- Tables  
d'abondance  
taxonomiques et  
fonctionnelles

**6**

# 3. Analyses méta-omiques

## quelques rappels de vocabulaire

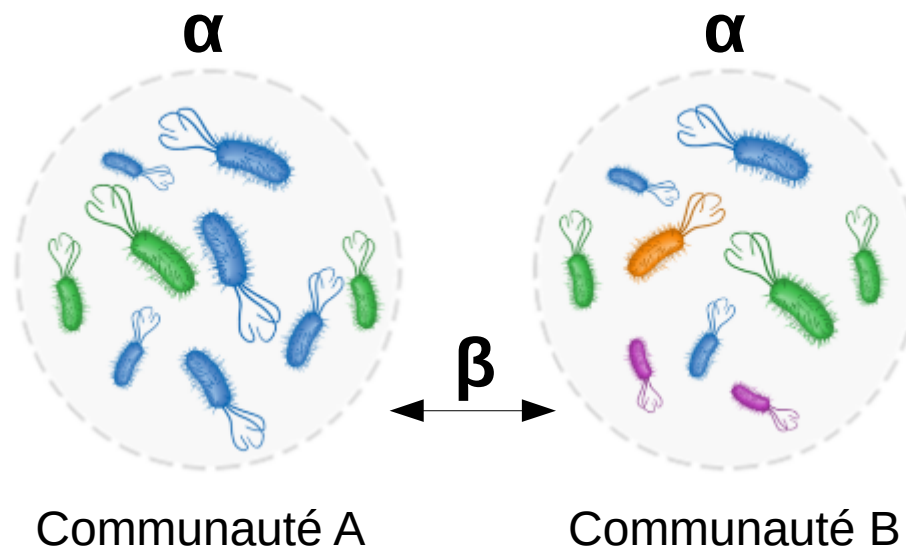


1. Introduction
2. Applications
3. Analyses méta-omiques
  - 3.1. La méta-génétique
  - 3.2. La méta-génomique
- 4. Analyses exploratoires**
5. Impact carbone du calcul



# 4. Analyses exploratoires

## Analyse de la biodiversité

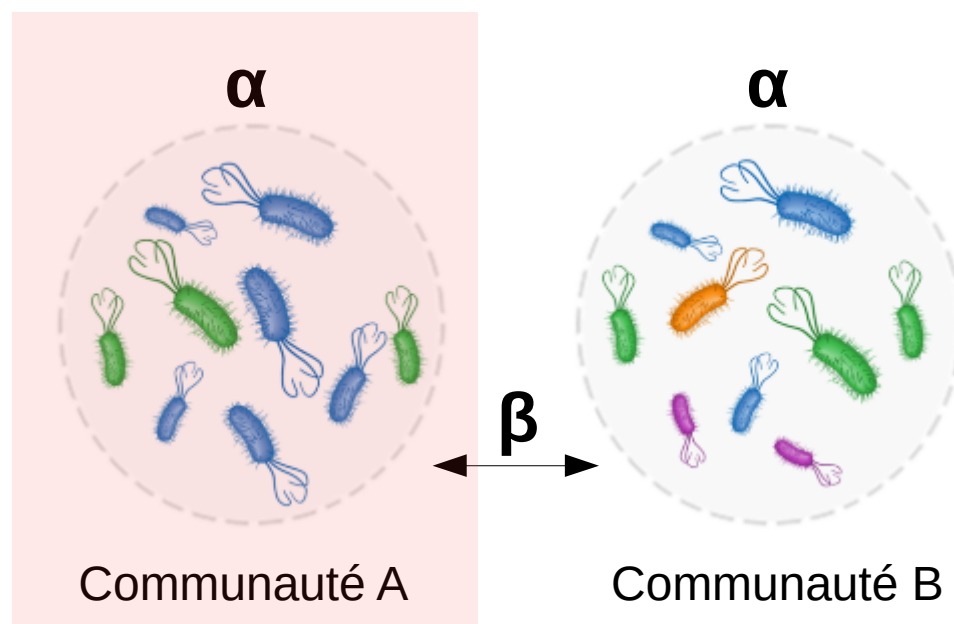


Diversité  $\alpha$  = diversité au sein d'une communauté.

Diversité  $\beta$  = diversité entre communautés.

# 4. Analyses exploratoires

## Analyse de la biodiversité



Diversité  $\alpha$  = diversité au sein d'une communauté.

Diversité  $\beta$  = diversité entre communautés.

## 4. Analyses exploratoires

### Analyse de la biodiversité – diversité $\alpha$

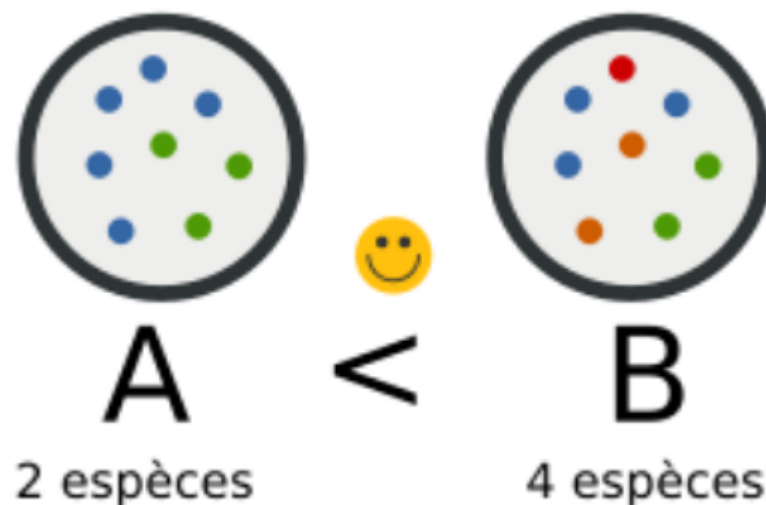
**Richesse** : nombre d'OTUs ou groupe fonctionnel au sein d'une communauté.  
Caractérise la composition.



## 4. Analyses exploratoires

### Analyse de la biodiversité – diversité $\alpha$

**Richesse** : nombre d'OTUs ou groupe fonctionnel au sein d'une communauté.  
Caractérise la composition.



**B est plus diversifié que A** car il contient deux fois plus d'espèces.



## 4. Analyses exploratoires

### Analyse de la biodiversité – diversité $\alpha$

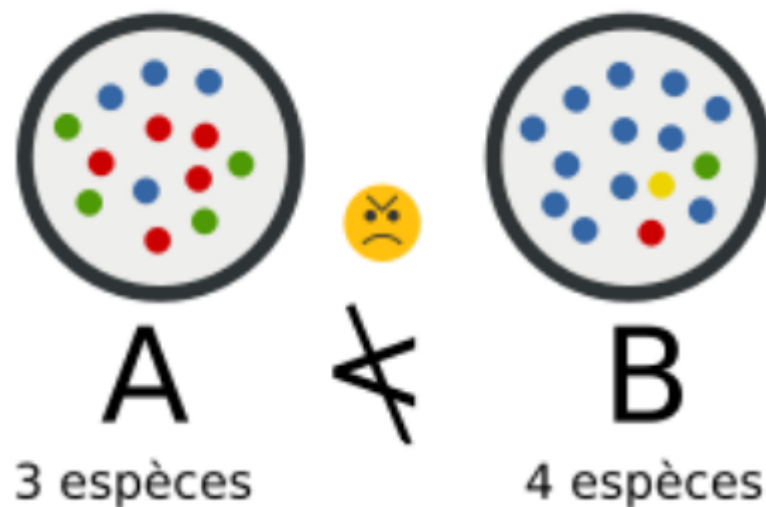
**Richesse** : nombre d'OTUs ou groupe fonctionnel au sein d'une communauté.  
Caractérise la composition.



## 4. Analyses exploratoires

### Analyse de la biodiversité – diversité $\alpha$

**Richesse** : nombre d'OTUs ou groupe fonctionnel au sein d'une communauté.  
Caractérise la composition.



**B** contient plus d'espèce mais **semble moins diversifié !**

## 4. Analyses exploratoires

### Analyse de la biodiversité – diversité $\alpha$

**Indice de Shannon** : indice reflète aussi bien le nombre d'espèces que leurs abondances.

$$H(X) = H_2(X) = - \sum_{i=1}^n P_i \log_2 P_i.$$

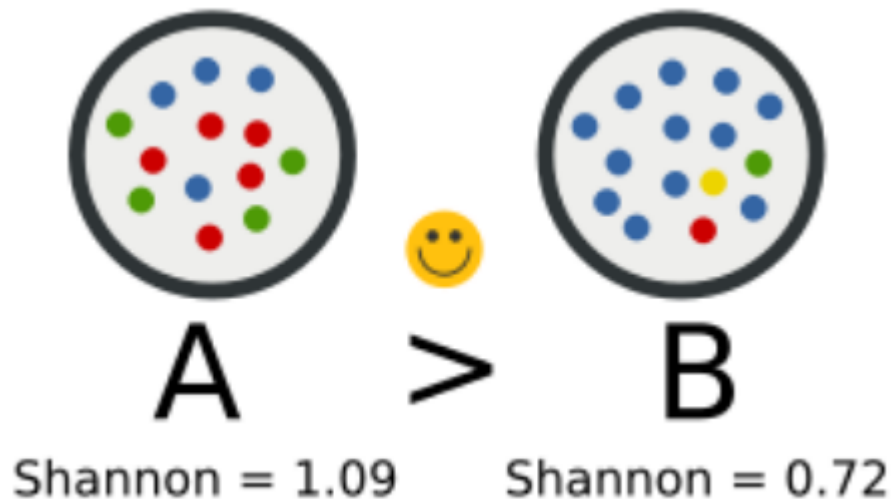


# 4. Analyses exploratoires

## Analyse de la biodiversité – diversité $\alpha$

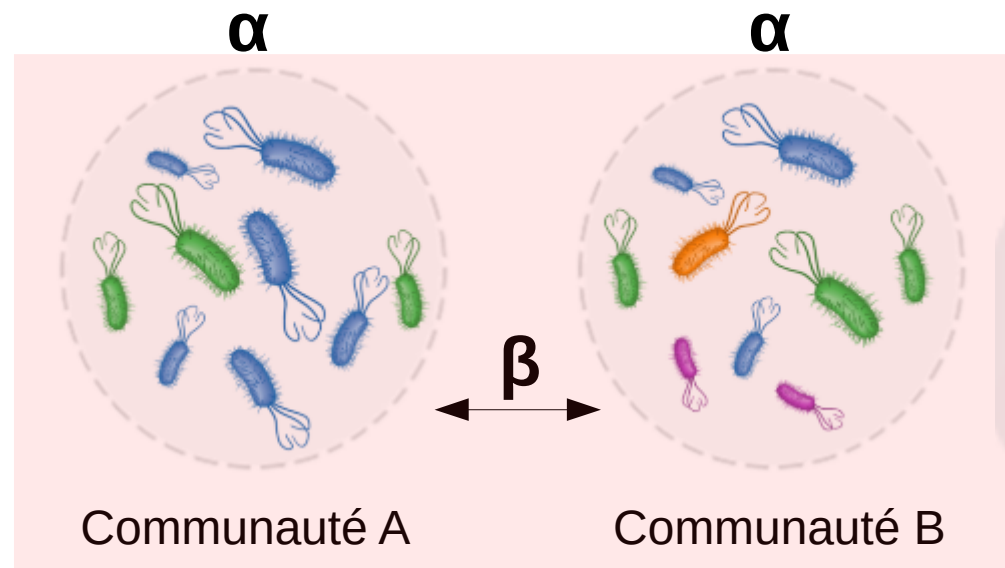
**Indice de Shannon** : indice reflète aussi bien le nombre d'espèce que leurs abondances.

$$H(X) = H_2(X) = - \sum_{i=1}^n P_i \log_2 P_i.$$



# 4. Analyses exploratoires

## Analyse de la biodiversité



Diversité  $\alpha$  = diversité au sein d'une communauté.

Diversité  $\beta$  = diversité entre communautés.

# 4. Analyses exploratoires

## Analyse de la biodiversité – diversité $\beta$

### Propriétés :

- échelonnées entre 0 et 1,
- 2 échantillons identiques :  $\text{Dist}(A,A) = 0$
- Si aucun OTUs partagé entre 2 échantillon :  $\text{Dist}(A,B) = 1$

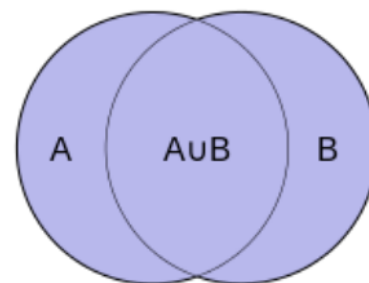
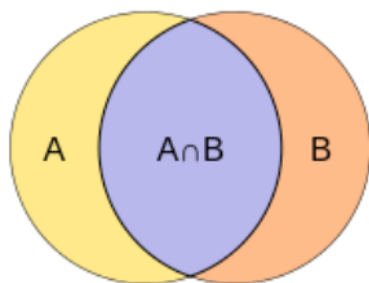
	Categorical	Phylogenetic
Presence/ Absence	Jaccard	Unifrac
Quantitative Abundance	Bray-Curtis	Weighted Unifrac

# 4. Analyses exploratoires

## Analyse de la biodiversité – diversité $\beta$

**Jaccard** : Fraction d'espèce spécifique à la communauté A ou B.

$$\begin{aligned}\text{Dist}(A, B) &= 1 - (A \cap B)/(A \cup B) \\ &= ((\mathbf{x}_A > 0) \& (\mathbf{x}_B > 0))/((\mathbf{x}_A > 0) \mid (\mathbf{x}_B > 0))\end{aligned}$$

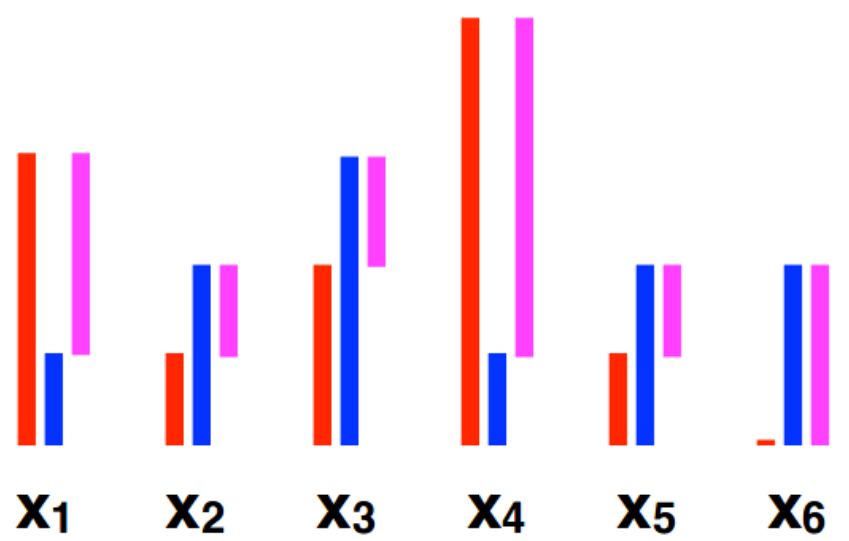


# 4. Analyses exploratoires

## Analyse de la biodiversité – diversité $\beta$

**Bray-Curtis** : Fraction de la communauté spécifique à la communauté A ou B.

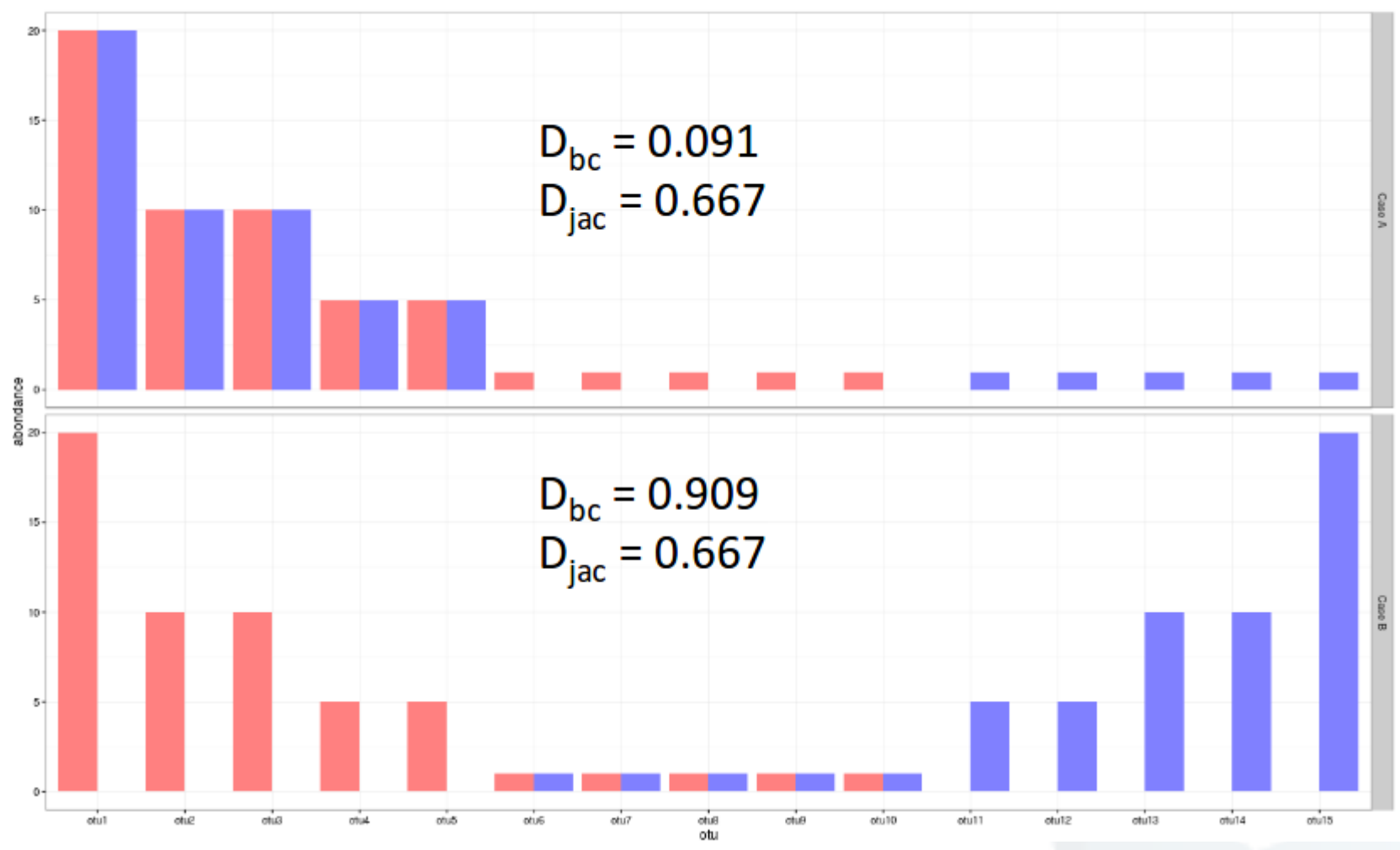
$$\text{Dist}(x, y) = \frac{\sum |x_i - y_i|}{\sum x_i + \sum y_i} = \frac{\text{pink}}{\text{orange} + \text{blue}}$$





# 4. Analyses exploratoires

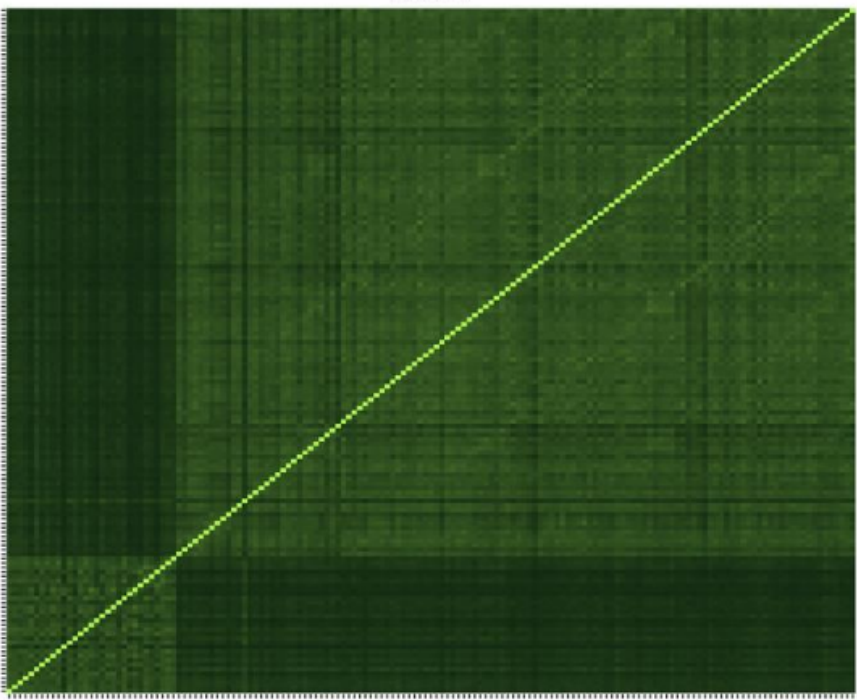
## Analyse de la biodiversité – diversité $\beta$



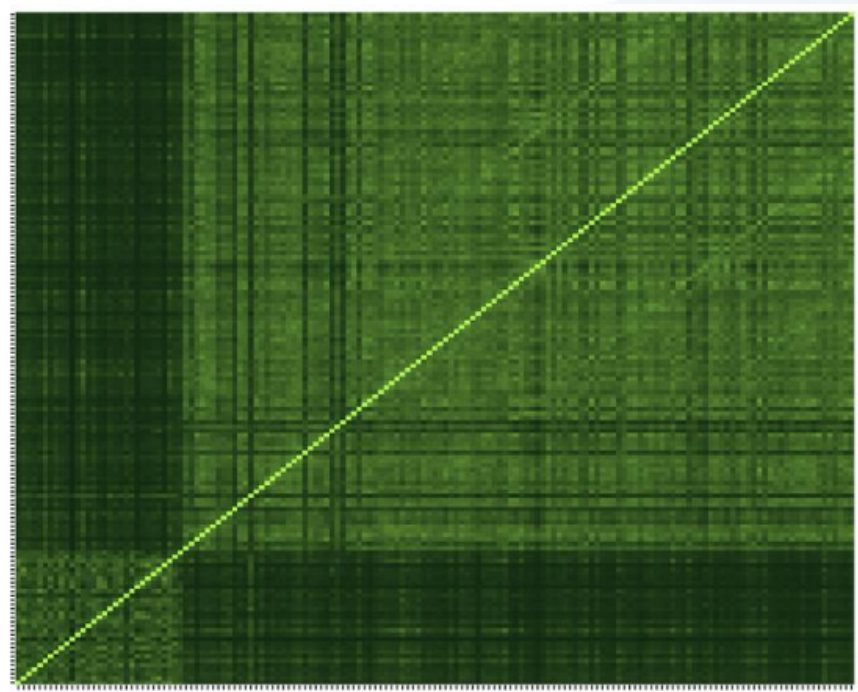
# 4. Analyses exploratoires

## Analyse de la biodiversité – diversité $\beta$

Jaccard



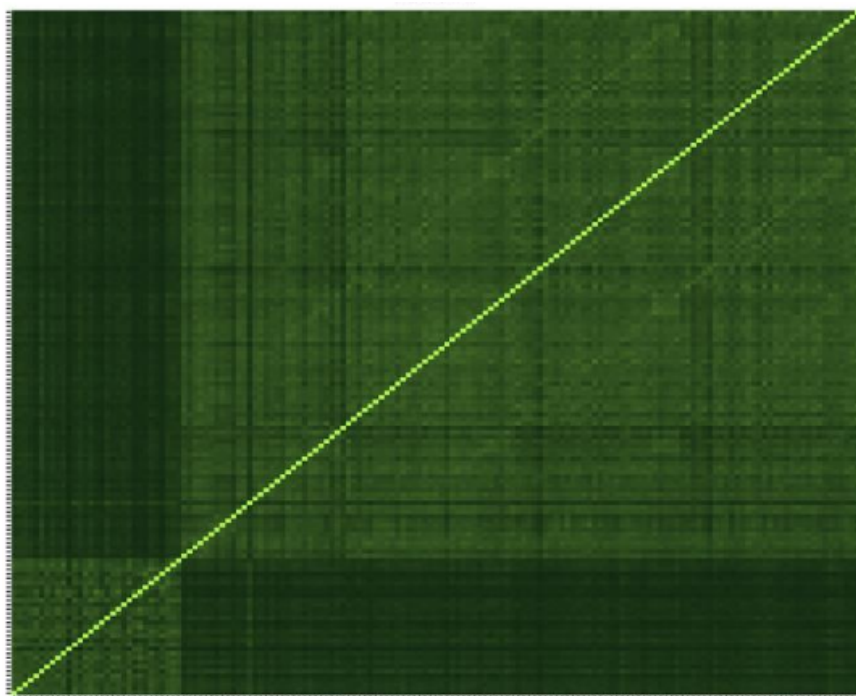
Bray-Curtis



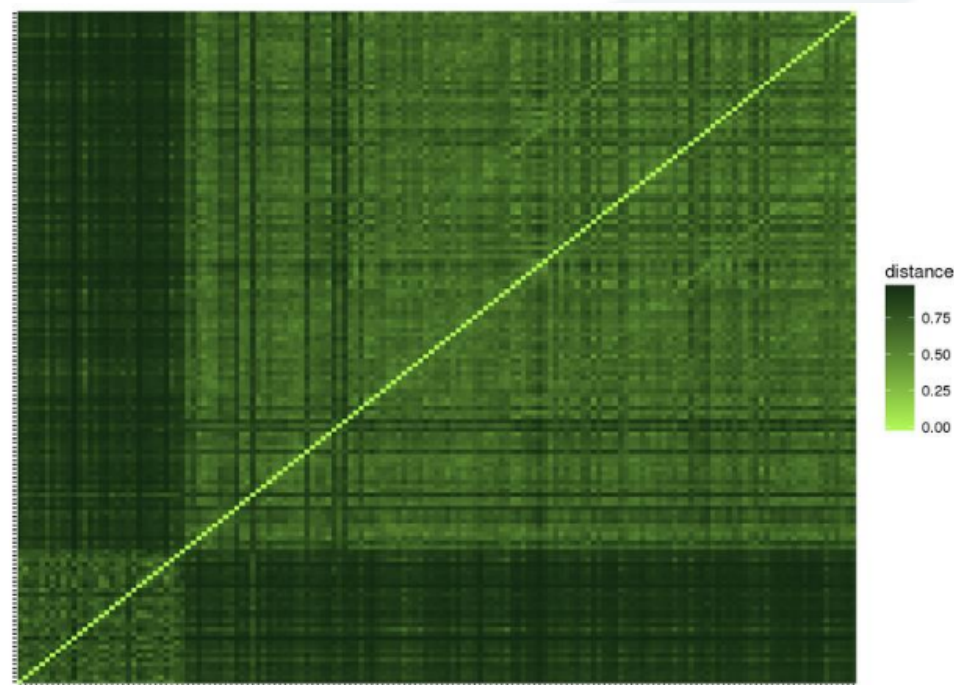
# 4. Analyses exploratoires

## Analyse de la biodiversité – diversité $\beta$

Jaccard



Bray-Curtis



**Jaccard > Bray-Curtis** : les OTUs abondants sont partagés

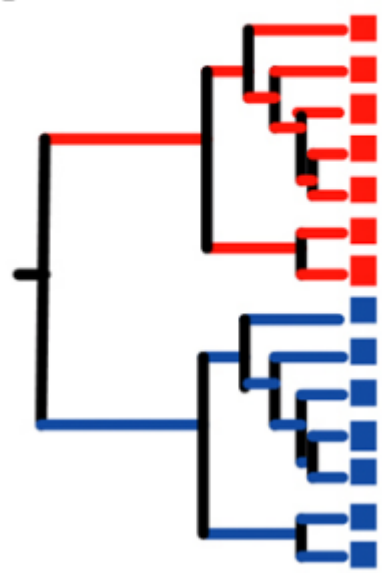
# 4. Analyses exploratoires

## Analyse de la biodiversité – diversité $\beta$

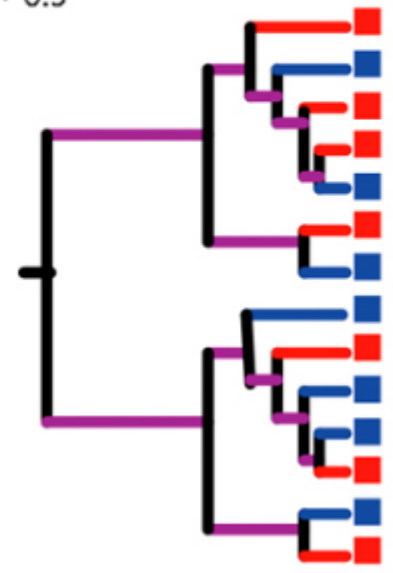
**Unifrac** : Fraction de l'arbre spécifique à la communauté A ou B.

$$\text{Dist}(x, y) = \frac{\text{red} + \text{blue}}{\text{red} + \text{blue} + \text{purple}}$$

D = 1



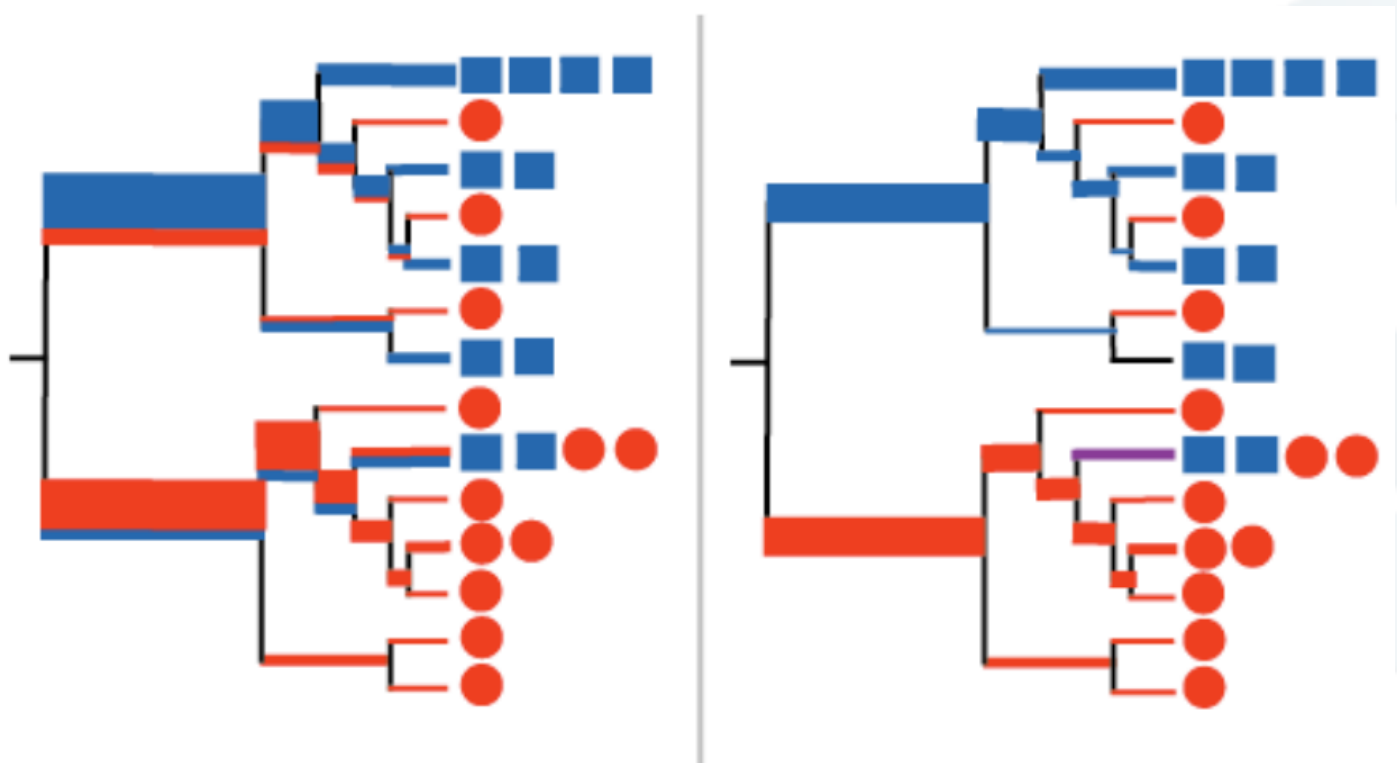
D = ~ 0.5



# 4. Analyses exploratoires

## Analyse de la biodiversité – diversité $\beta$

**Weighted Unifrac** : Fraction de la diversité spécifique à la communauté A ou B.

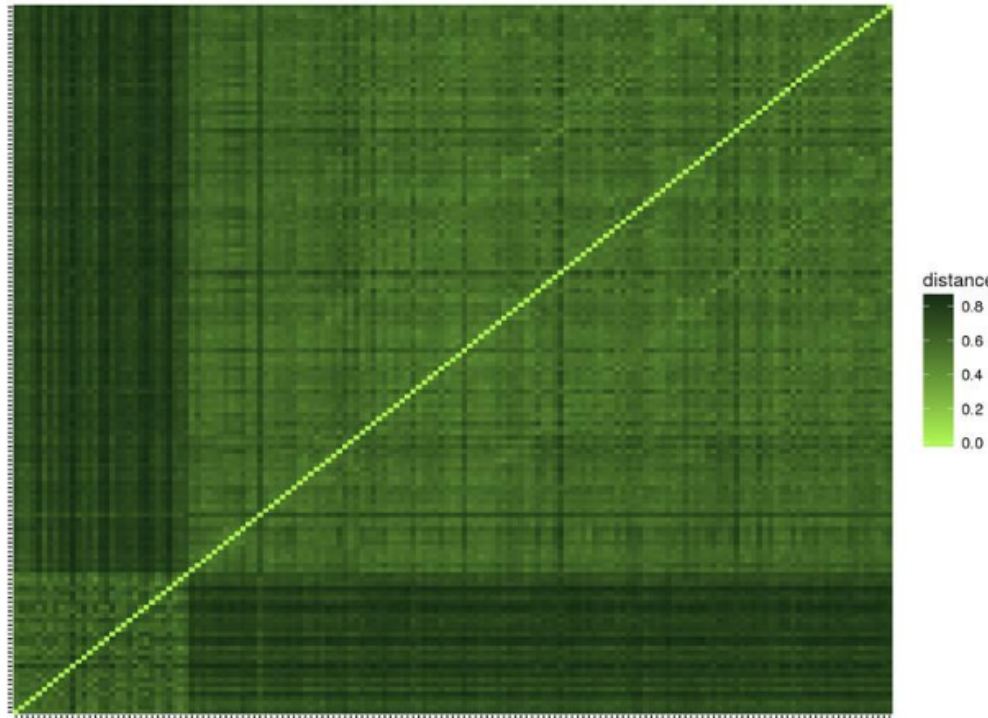




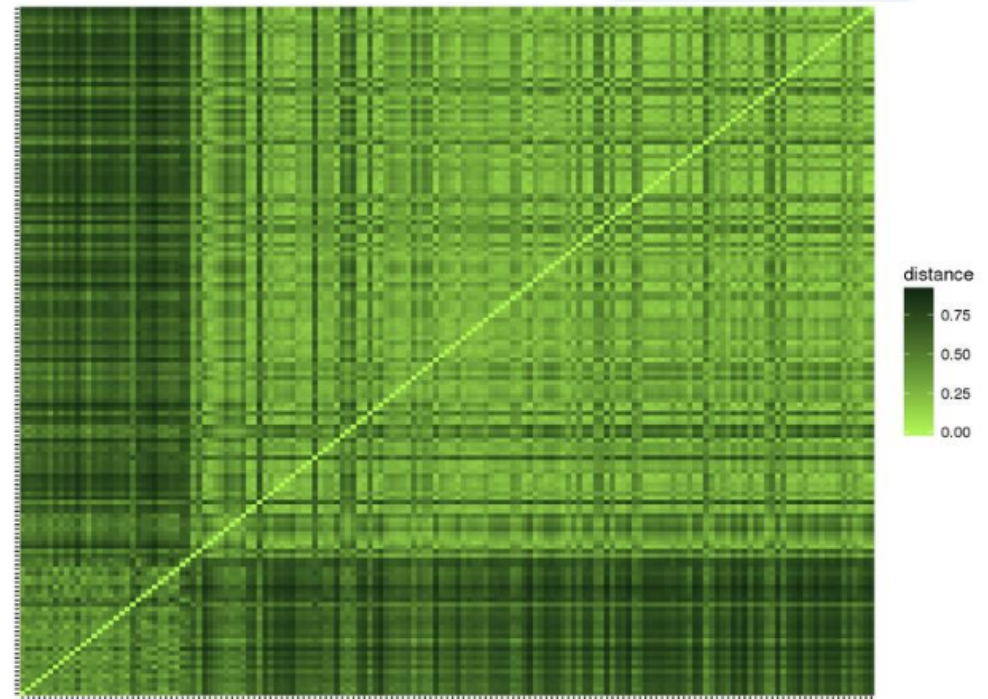
# 4. Analyses exploratoires

## Analyse de la biodiversité – diversité $\beta$

UniFrac



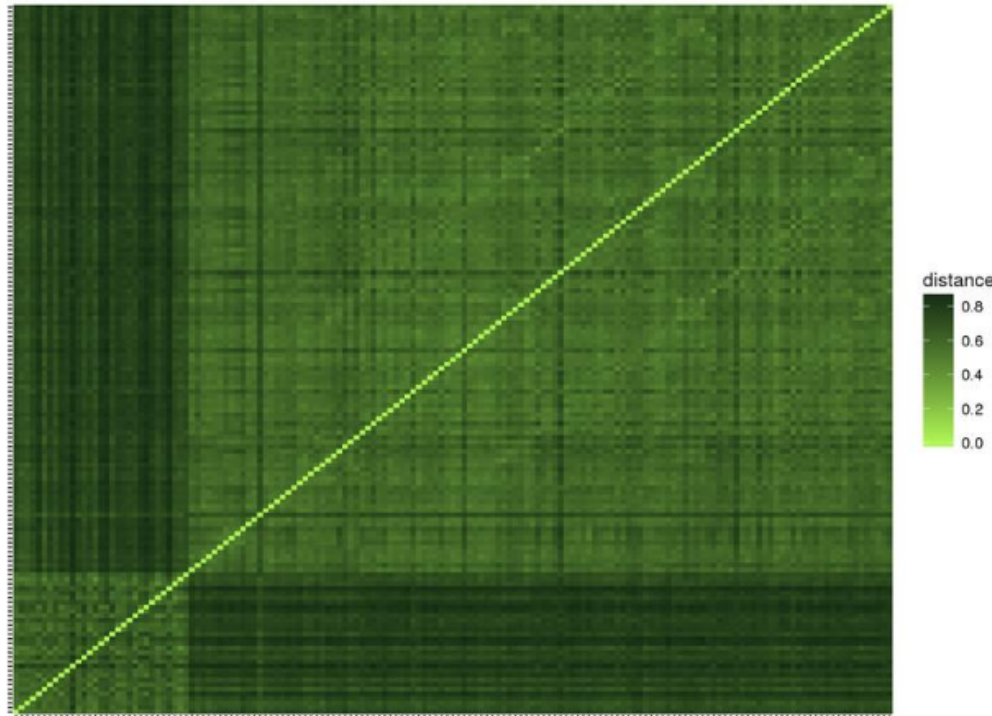
wUniFrac



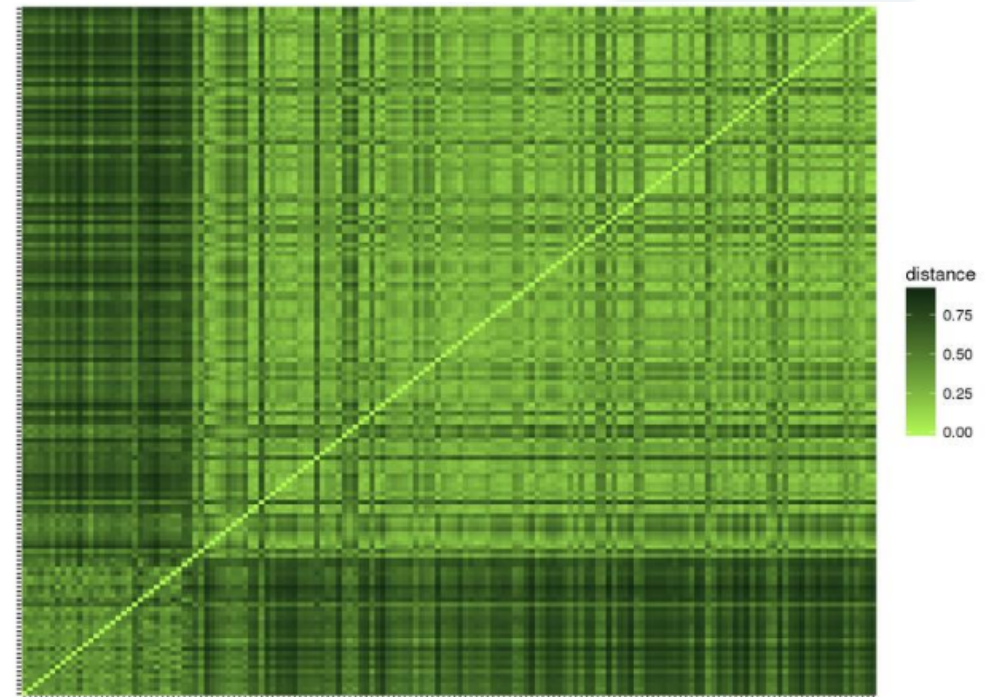
# 4. Analyses exploratoires

## Analyse de la biodiversité – diversité $\beta$

UniFrac



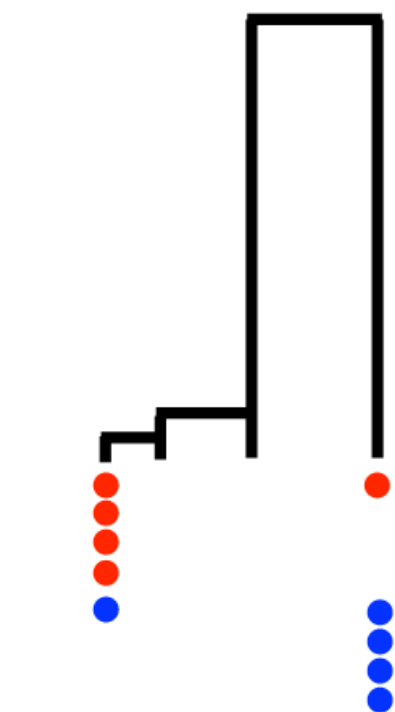
wUniFrac



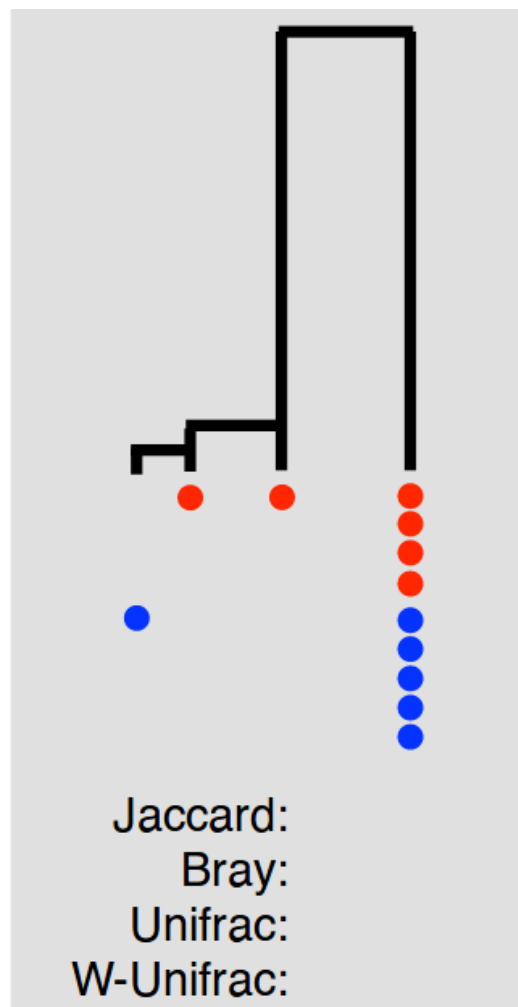
**UniFrac > wUniFrac** : les OTUs abondants sont proches phylogénétiquement

# 4. Analyses exploratoires

## Analyse de la biodiversité – diversité $\beta$



Jaccard:  
Bray:  
Unifrac:  
W-Unifrac:



Jaccard:  
Bray:  
Unifrac:  
W-Unifrac:

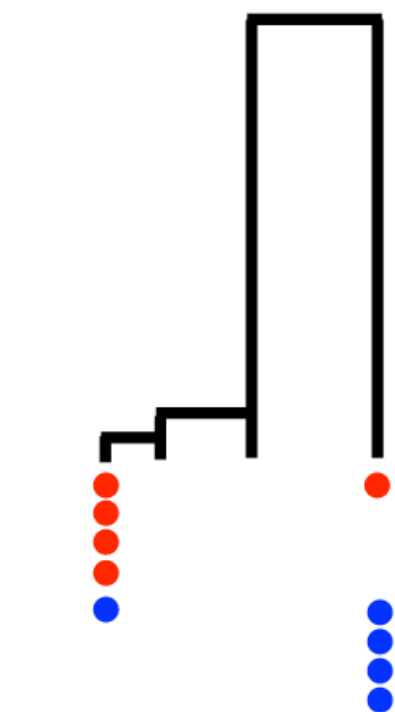


Jaccard:  
Bray:  
Unifrac:  
W-Unifrac:

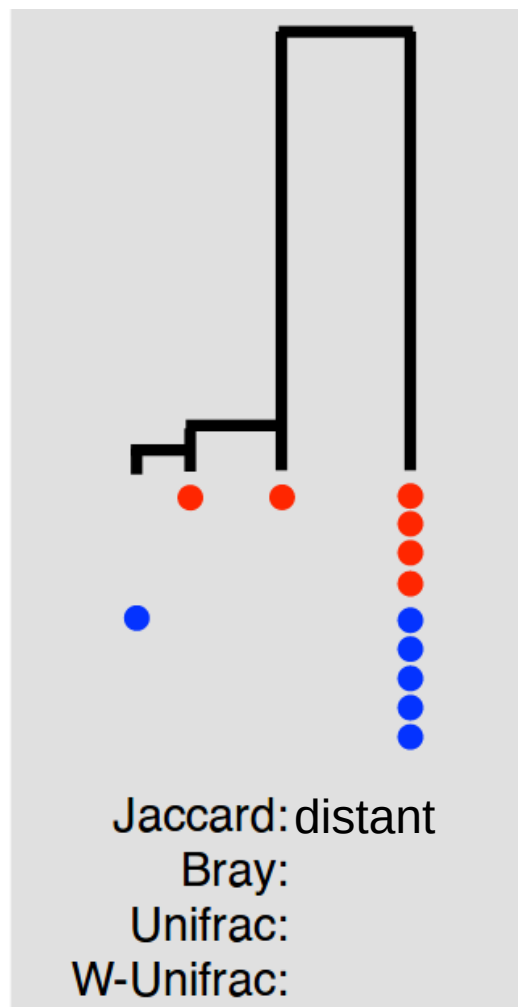


# 4. Analyses exploratoires

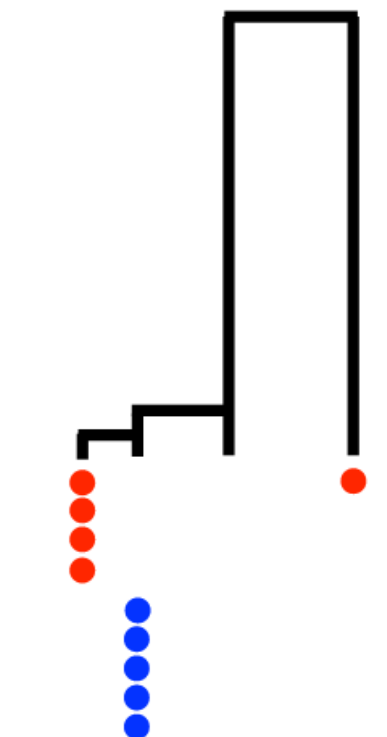
## Analyse de la biodiversité – diversité $\beta$



Jaccard:  $d=0$   
Bray:  
Unifrac:  
W-Unifrac:



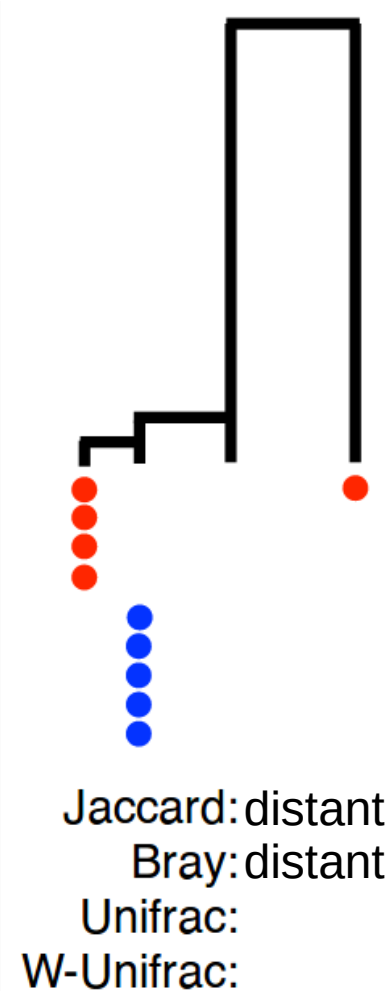
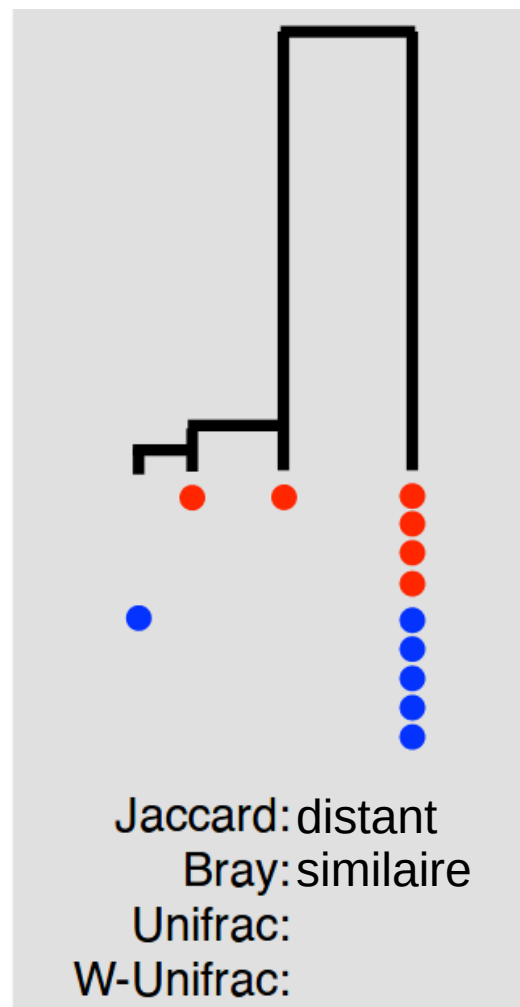
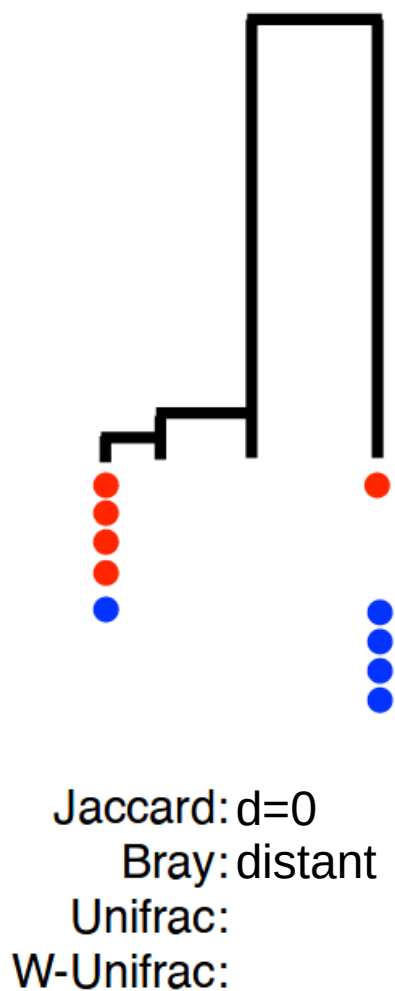
Jaccard: distant  
Bray:  
Unifrac:  
W-Unifrac:



Jaccard: distant  
Bray:  
Unifrac:  
W-Unifrac:

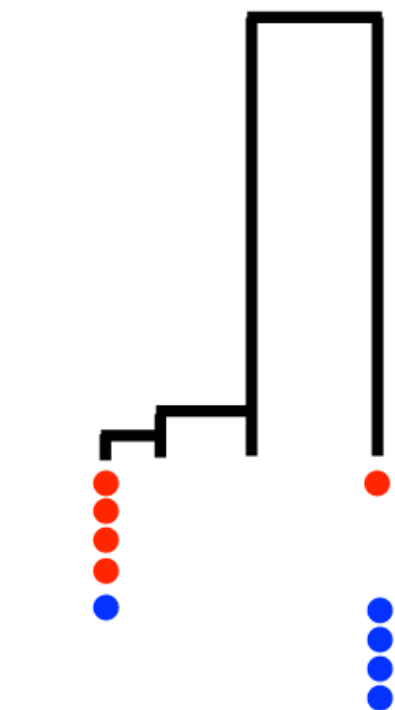
# 4. Analyses exploratoires

## Analyse de la biodiversité – diversité $\beta$

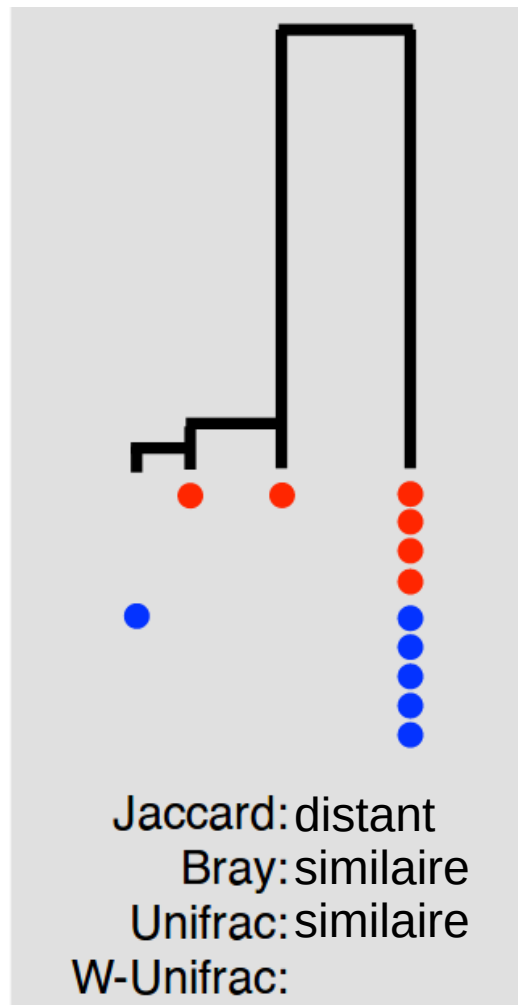


# 4. Analyses exploratoires

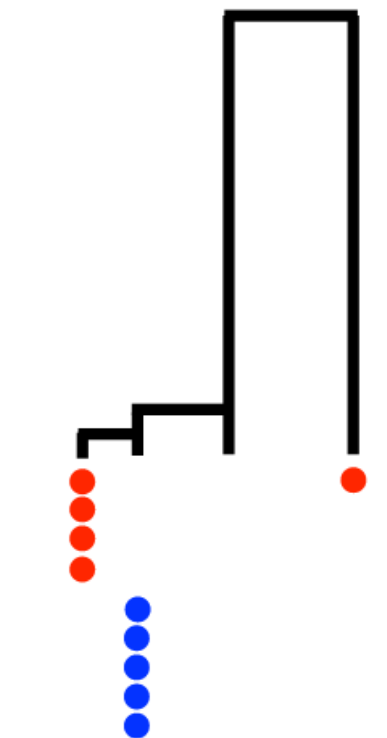
## Analyse de la biodiversité – diversité $\beta$



Jaccard:  $d=0$   
 Bray: distant  
 Unifrac:  $d=0$   
 W-Unifrac:



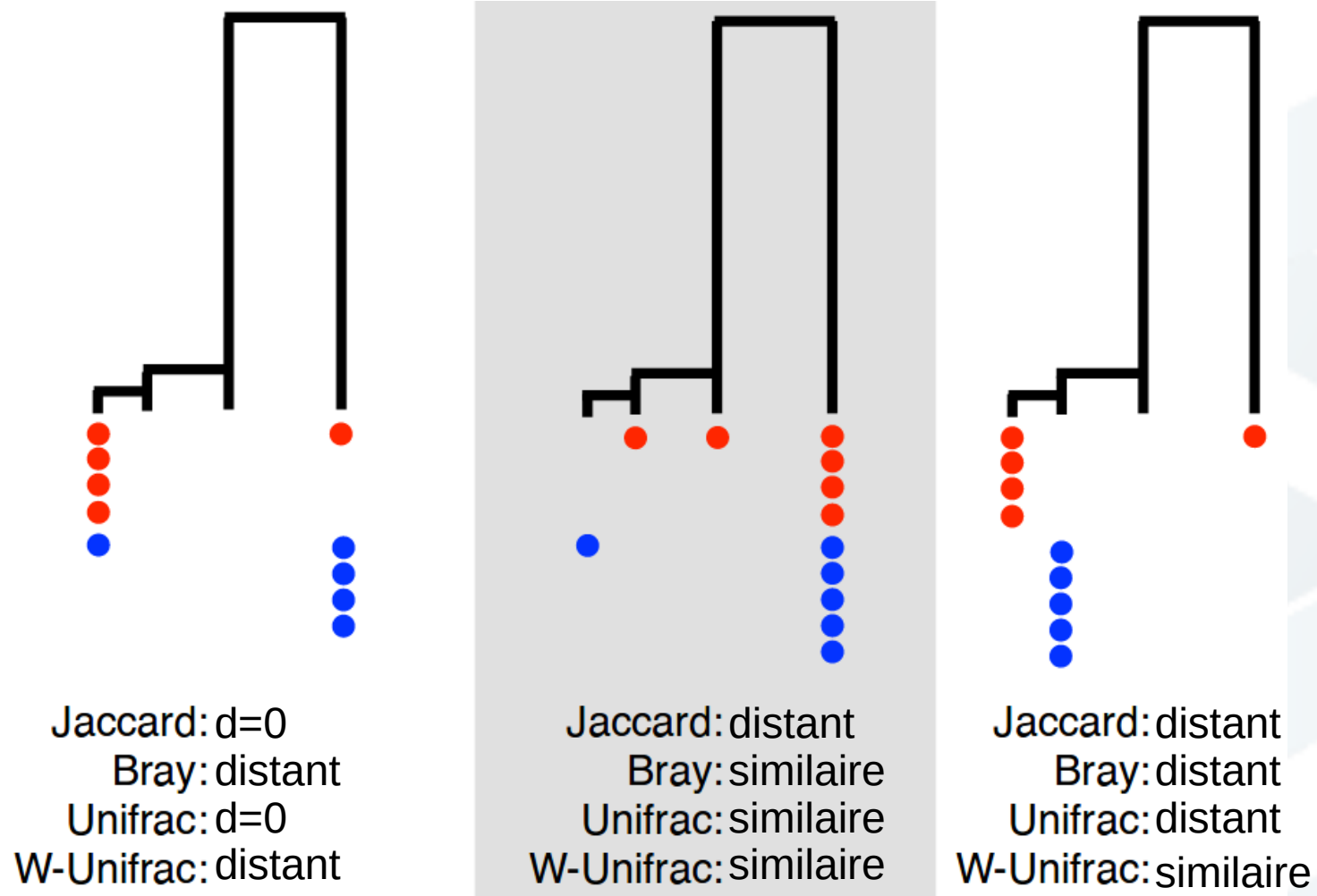
Jaccard: distant  
 Bray: similaire  
 Unifrac: similaire  
 W-Unifrac:



Jaccard: distant  
 Bray: distant  
 Unifrac: distant  
 W-Unifrac:

# 4. Analyses exploratoires

## Analyse de la biodiversité – diversité $\beta$



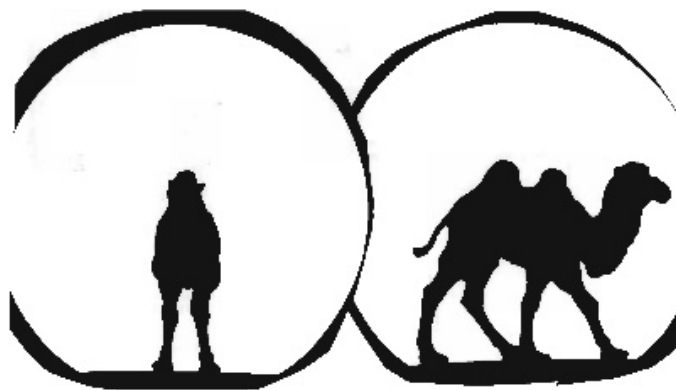
# 4. Analyses exploratoires

## Méthodes exploratoires - MDS

**Objectif** : Projeter des données de grande dimension dans un sous espace de plus faible dimension.

L'**ACP** (Analyse en Composantes Principales) recherche des combinaisons linéaires d'OTUs qui

- sont non corrélés,
- préservent au mieux la variance de la composition des communautés.



Chameau ou dromadaire ?

Mais les données méta-omics :

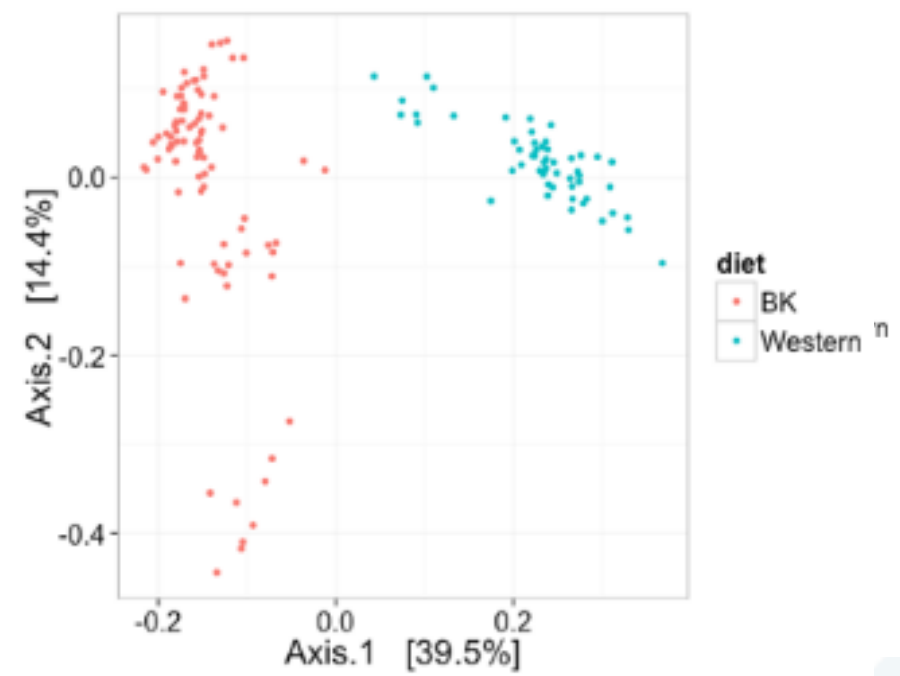
- peuvent être corrélées,
- l'ACP n'est pas adaptée pour capturer la diversité.

# 4. Analyses exploratoires

## Méthodes exploratoires - MDS

La **MDS** (Multi-Dimensional Scaling) est équivalente à l'ACP mais capture la diversité  $\beta$ .

N samples	P taxa
	0,1,5,1,0,1,2,1,0,0,9,...
	7,2,0,0,0,0,0,0,1,0,0,...
	0,0,0,0,0,0,8,0,0,0,1,...
	0,0,0,1,0,1,2,0,0,0,5,...
	0,1,0,2,0,0,0,1,0,0,4,...
	0,0,0,1,9,1,2,5,2,0,1,...
	0,0,0,0,0,1,2,1,8,0,0,...
	0,0,0,0,9,4,0,0,0,0,1,...



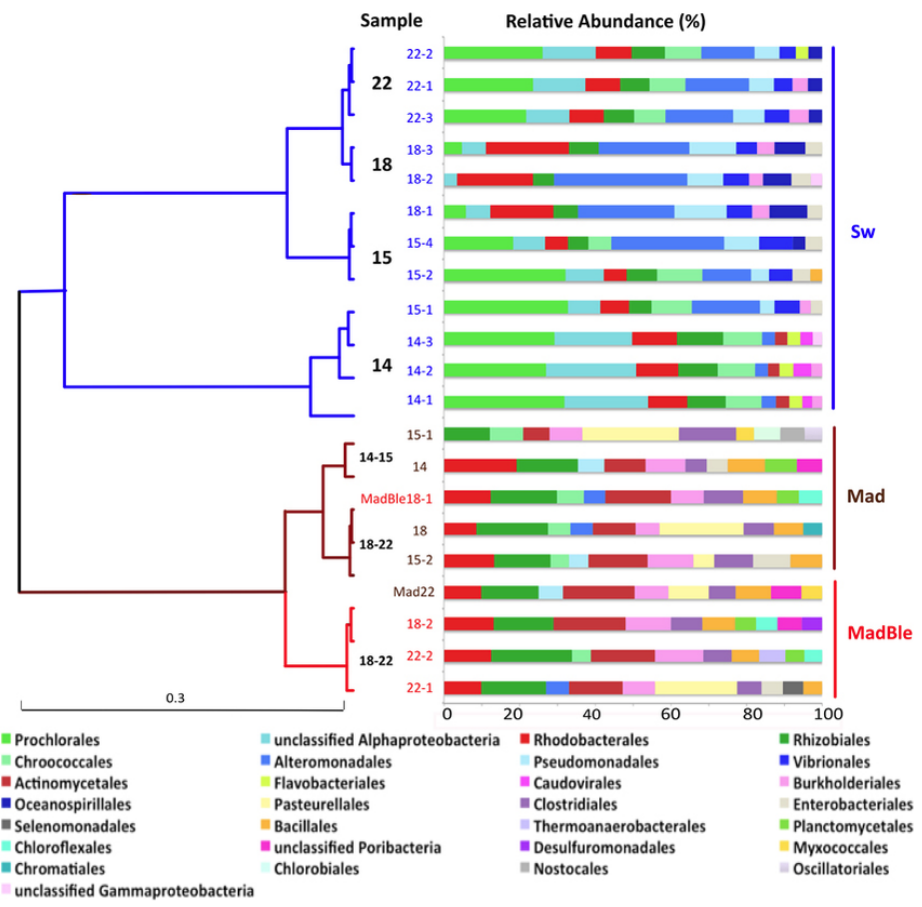
P-dimensions

2-dimensions

# 4. Analyses exploratoires

## Méthodes exploratoires – Classification hiérarchique

**Objectif :** Regrouper les échantillons selon leur diversité.



Sw = seawater  
Mad = corail sain  
MadBle = corail blanchi

# 4. Analyses exploratoires

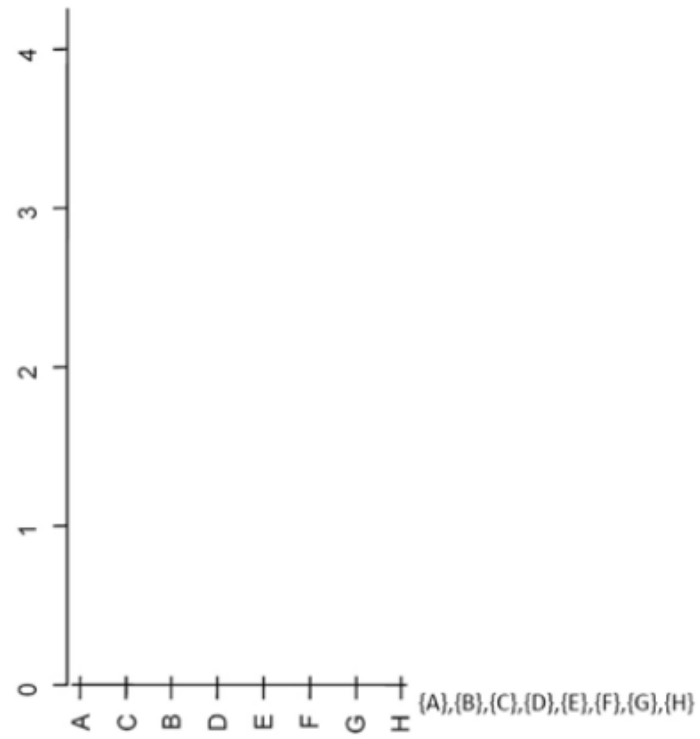
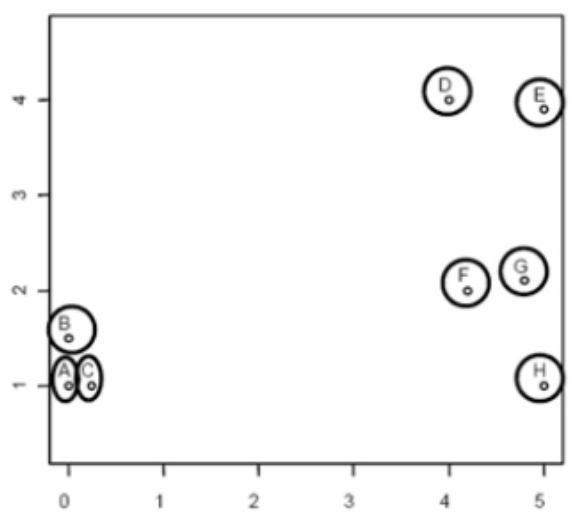
Méthodes exploratoires – Classification hiérarchique





# 4. Analyses exploratoires

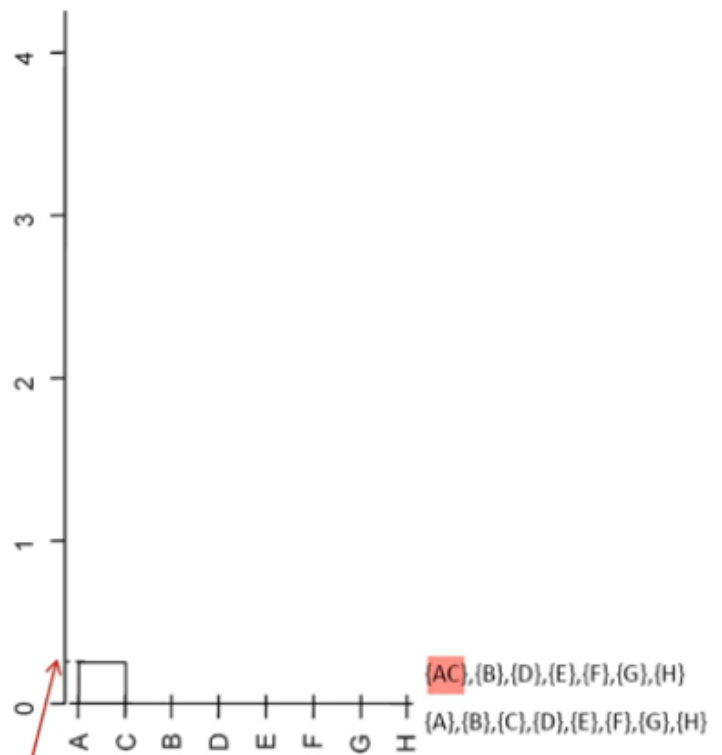
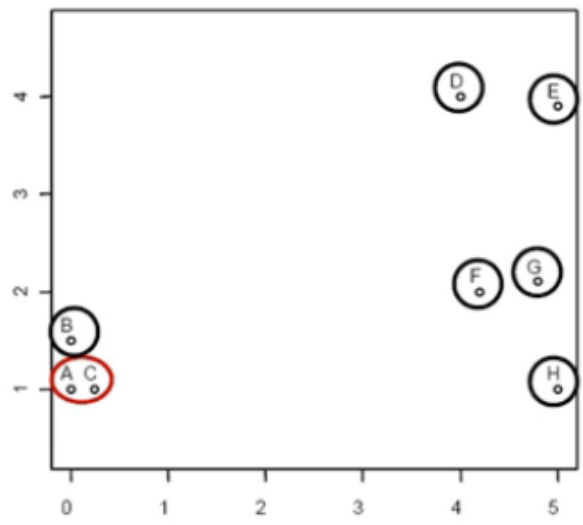
## Méthodes exploratoires – Classification hiérarchique



	A	B	C	D	E	F	G
B	0.50						
C	0.25	0.56					
D	5.00	4.72	4.80				
E	5.78	5.55	5.57	1.00			
F	4.32	4.23	4.07	2.01	2.06		
G	4.92	4.84	4.68	2.06	1.81	0.61	
H	5.00	5.02	4.75	3.16	2.90	1.28	1.12

# 4. Analyses exploratoires

## Méthodes exploratoires – Classification hiérarchique

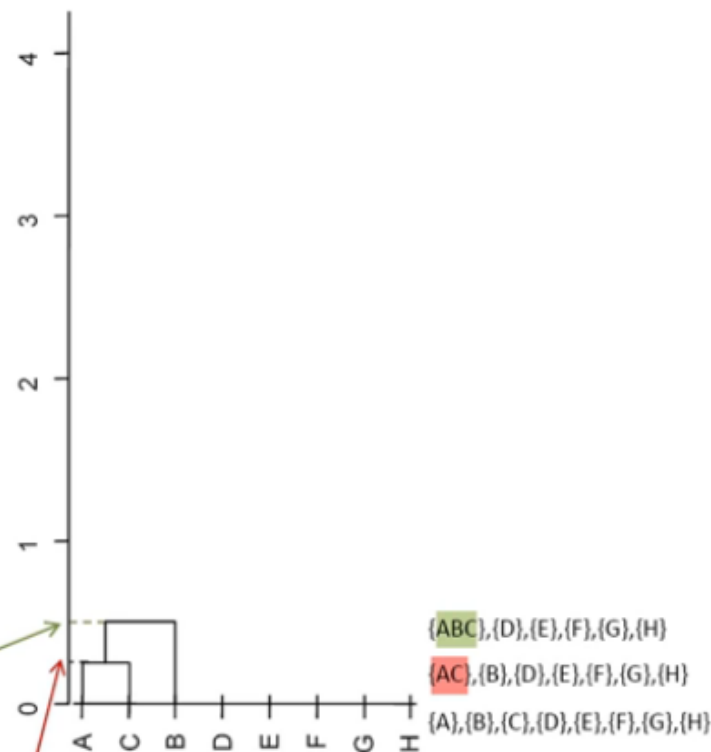
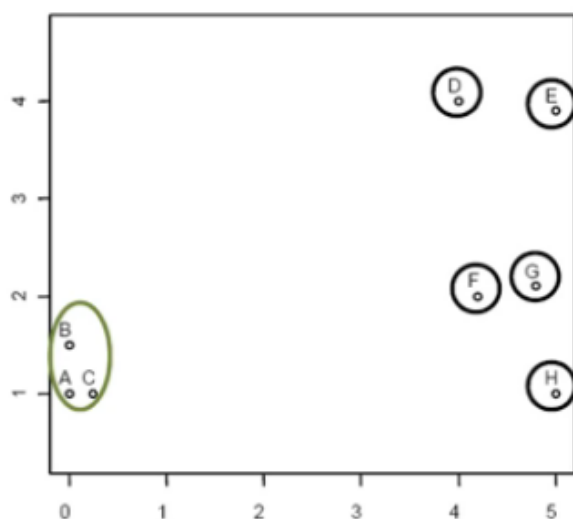


	A	B	C	D	E	F	G
B	0.50						
C	0.25	0.56					
D	5.00	4.72	4.80				
E	5.78	5.55	5.57	1.00			
F	4.32	4.23	4.07	2.01	2.06		
G	4.92	4.84	4.68	2.06	1.81	0.61	
H	5.00	5.02	4.75	3.16	2.90	1.28	1.12

1<sup>er</sup> regroupement

# 4. Analyses exploratoires

## Méthodes exploratoires – Classification hiérarchique



2<sup>e</sup> regroupement

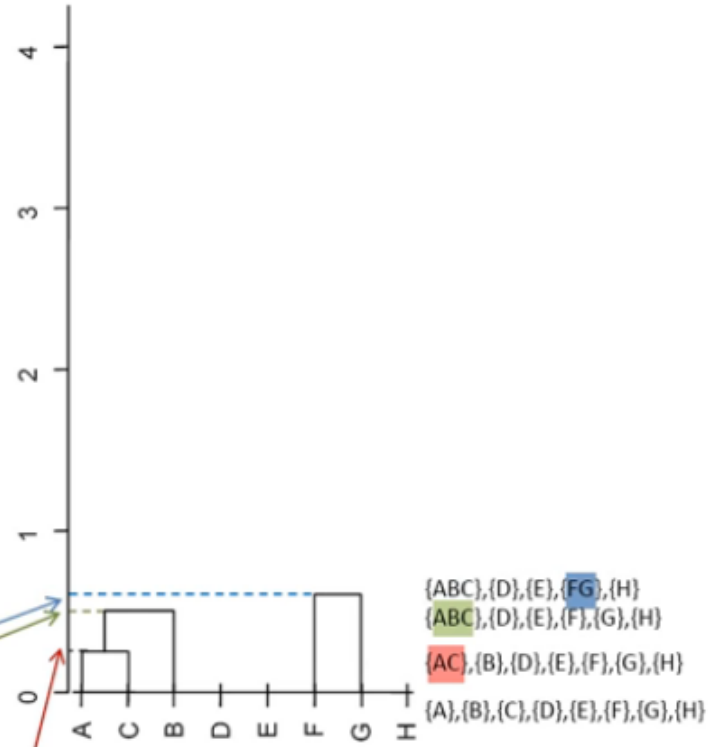
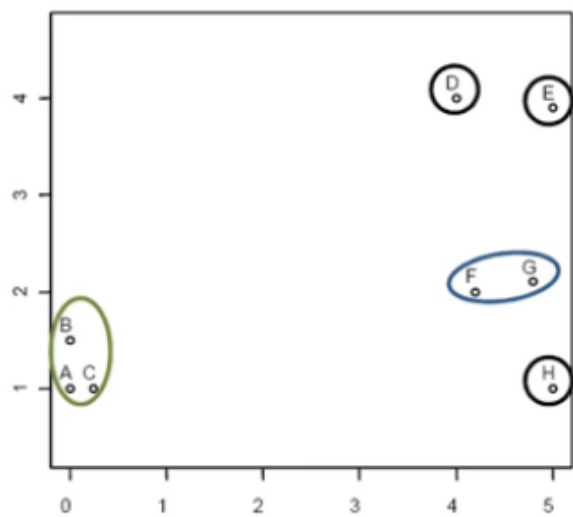
	AC	B	D	E	F	G
B	0.50					
D	4.80	4.72				
E	5.57	5.55	1.00			
F	4.07	4.23	2.01	2.06		
G	4.68	4.84	2.06	1.81	0.61	
H	4.75	5.02	3.16	2.90	1.28	1.12

	A	B	C	D	E	F	G
B	0.50						
C	0.25	0.56					
D	5.00	4.72	4.80				
E	5.78	5.55	5.57	1.00			
F	4.32	4.23	4.07	2.01	2.06		
G	4.92	4.84	4.68	2.06	1.81	0.61	
H	5.00	5.02	4.75	3.16	2.90	1.28	1.12

1<sup>er</sup> regroupement

# 4. Analyses exploratoires

## Méthodes exploratoires – Classification hiérarchique



3<sup>e</sup> regroupement

	ABC	D	E	F	G
D	4.72				
E	5.55	1.00			
F	4.07	2.01	2.06		
G	4.68	2.06	1.81	0.61	
H	4.75	3.16	2.90	1.28	1.12

2<sup>e</sup> regroupement

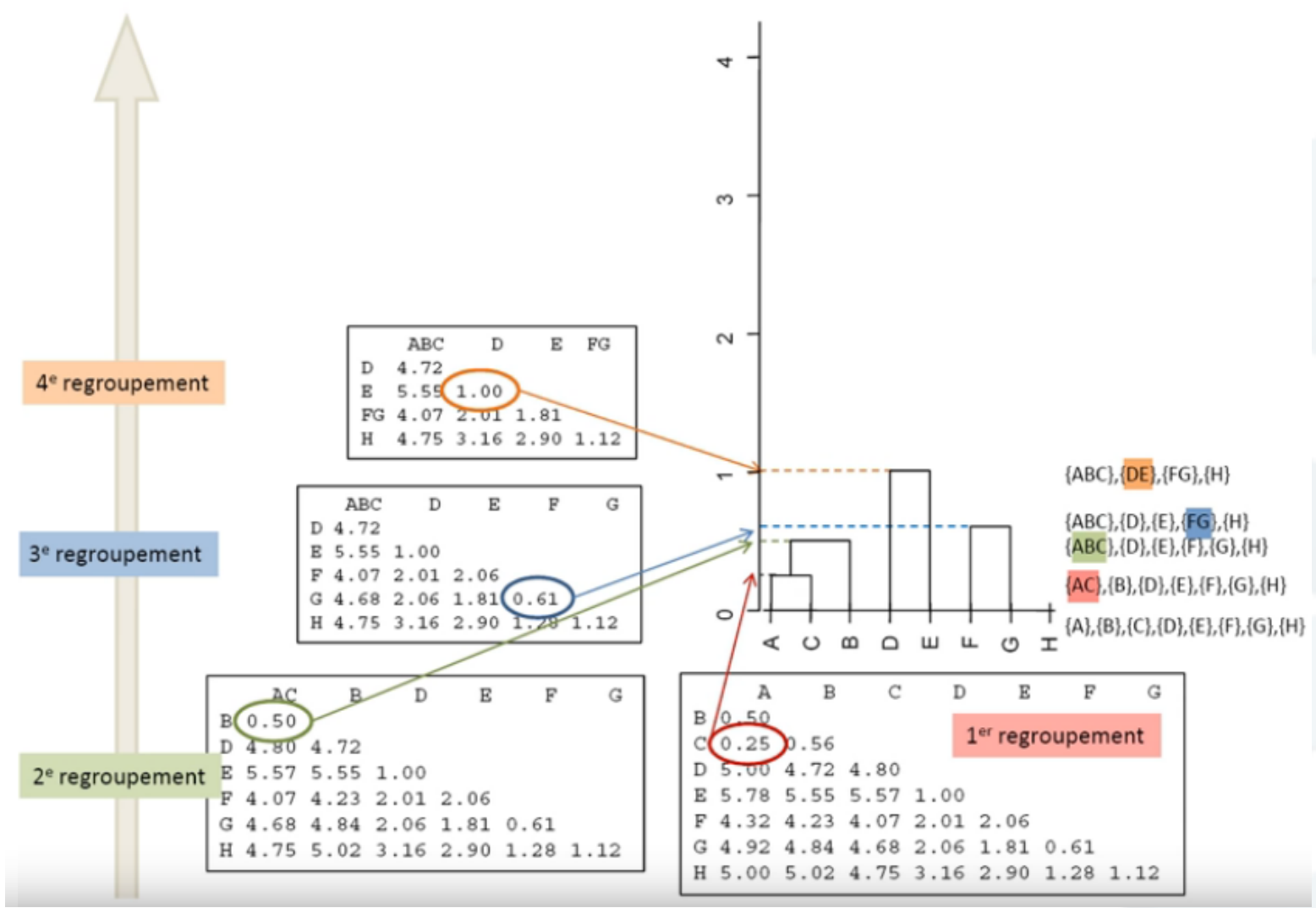
	AC	B	D	E	F	G
B	0.50					
D	4.80	4.72				
E	5.57	5.55	1.00			
F	4.07	4.23	2.01	2.06		
G	4.68	4.84	2.06	1.81	0.61	
H	4.75	5.02	3.16	2.90	1.28	1.12

1<sup>er</sup> regroupement

	A	B	C	D	E	F	G
B	0.50						
C	0.25	0.56					
D	5.00	4.72	4.80				
E	5.78	5.55	5.57	1.00			
F	4.32	4.23	4.07	2.01	2.06		
G	4.92	4.84	4.68	2.06	1.81	0.61	
H	5.00	5.02	4.75	3.16	2.90	1.28	1.12

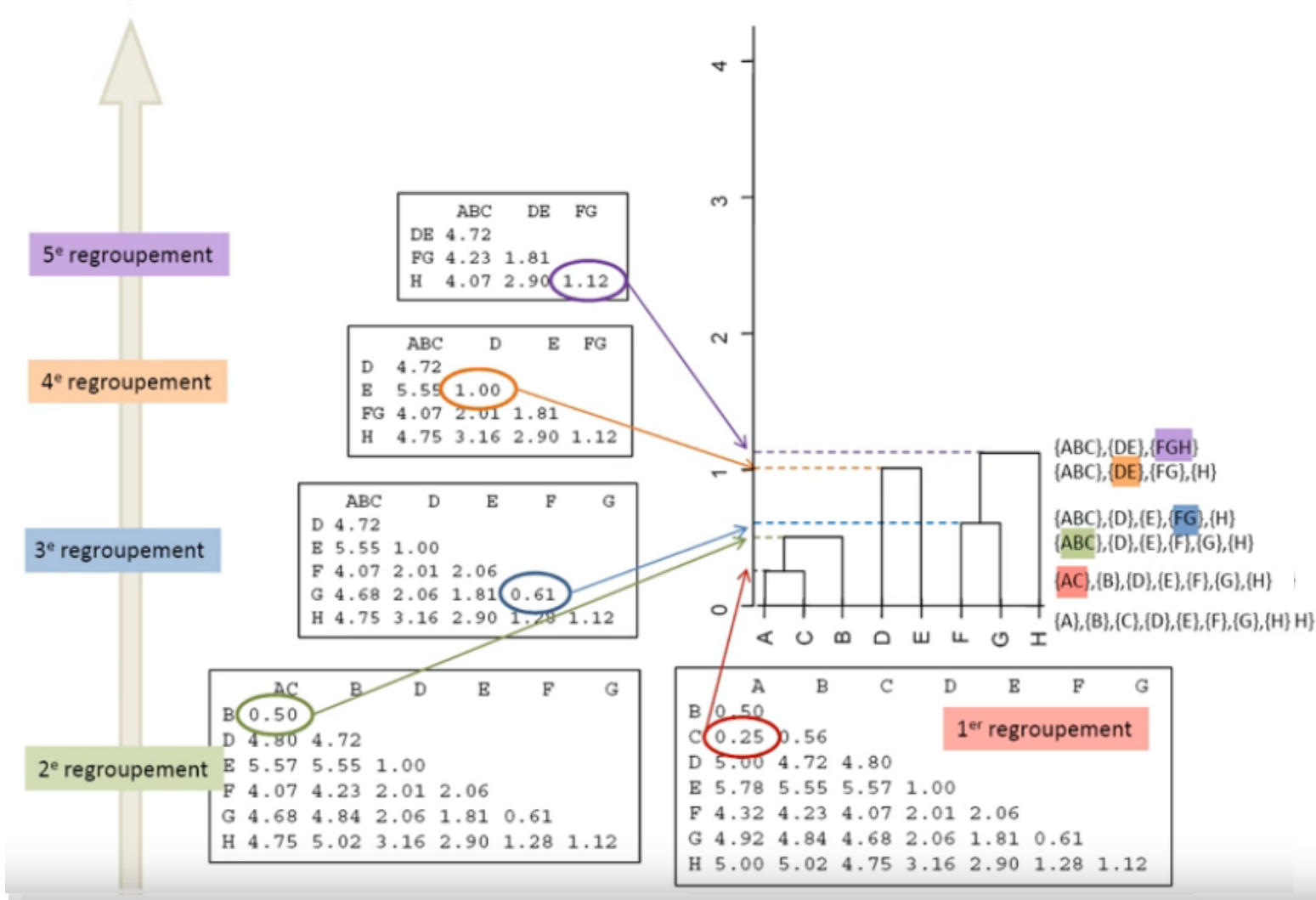
# 4. Analyses exploratoires

## Méthodes exploratoires – Classification hiérarchique



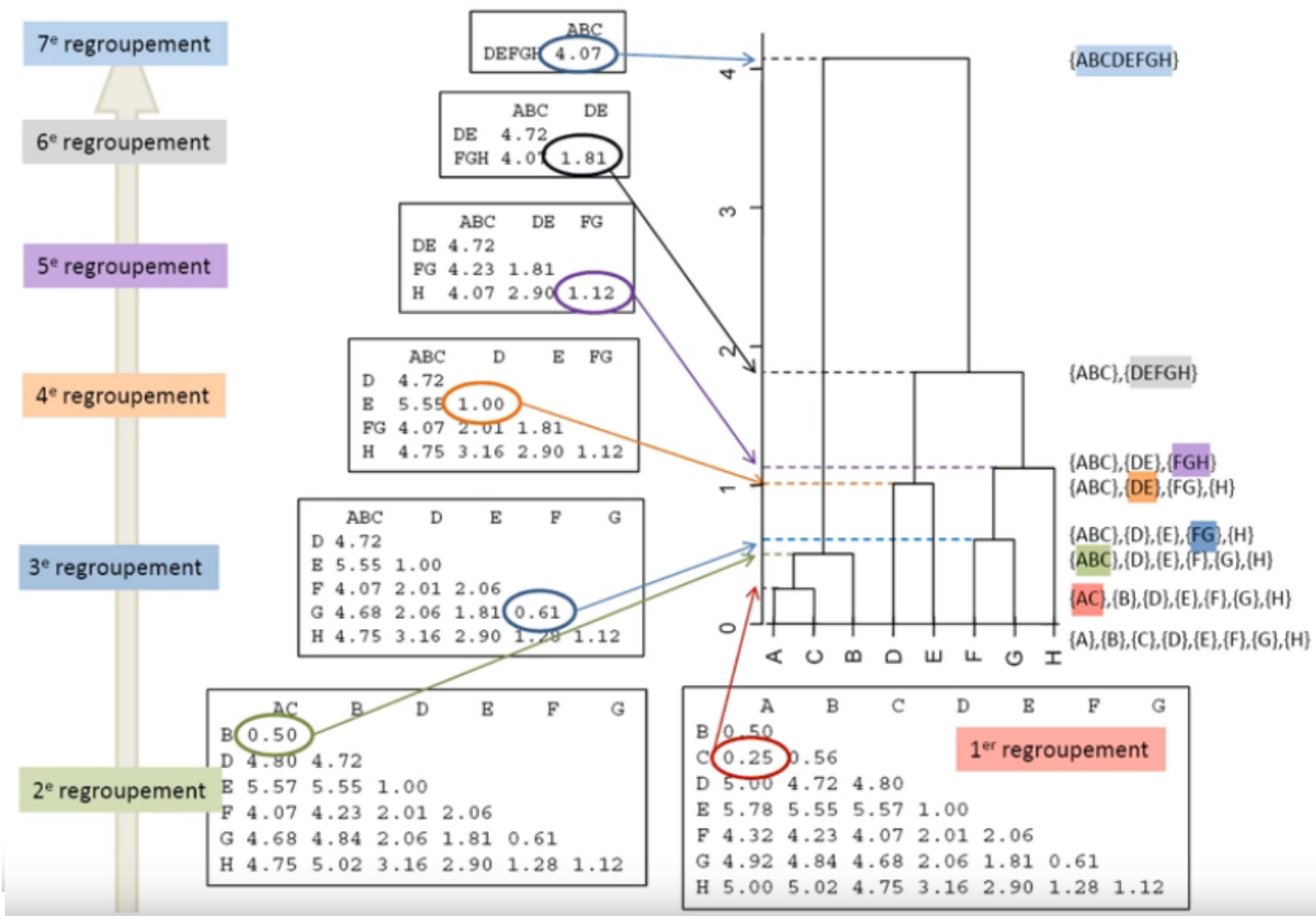
# 4. Analyses exploratoires

## Méthodes exploratoires – Classification hiérarchique



# 4. Analyses exploratoires

## Méthodes exploratoires – Classification hiérarchique



1. Introduction
2. Applications
3. Analyses mé $\alpha$ -omiques
  - 3.1. La mé $\alpha$ -géné $\alpha$ tique
  - 3.2. La mé $\alpha$ -génomique
4. Analyses exploratoires
- 5. Impact carbone du calcul**

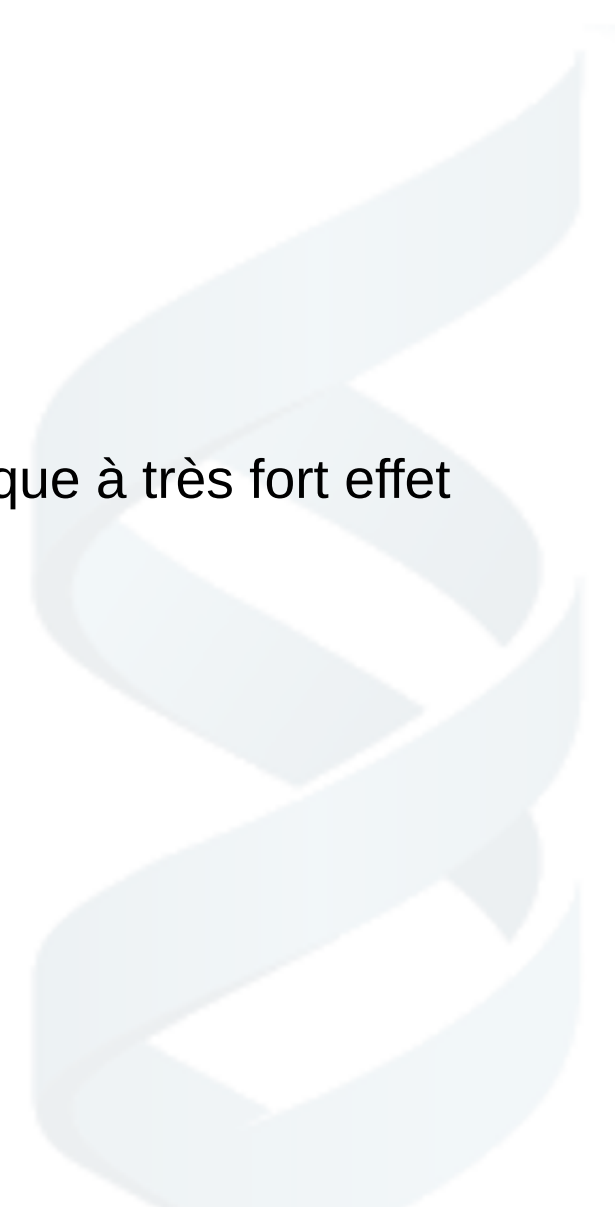




# 5. Impact carbone du calcul

## BGES du cluster GenoToul

- Matériel informatique : **fabrication, transport & eol**
- Consommation électrique
- **Fluide frigorigène** (fuites en R410A : fluide frigorigère à très fort effet de serre)
- **Déplacements** professionnel du personnel
- Déplacements domicile / travail

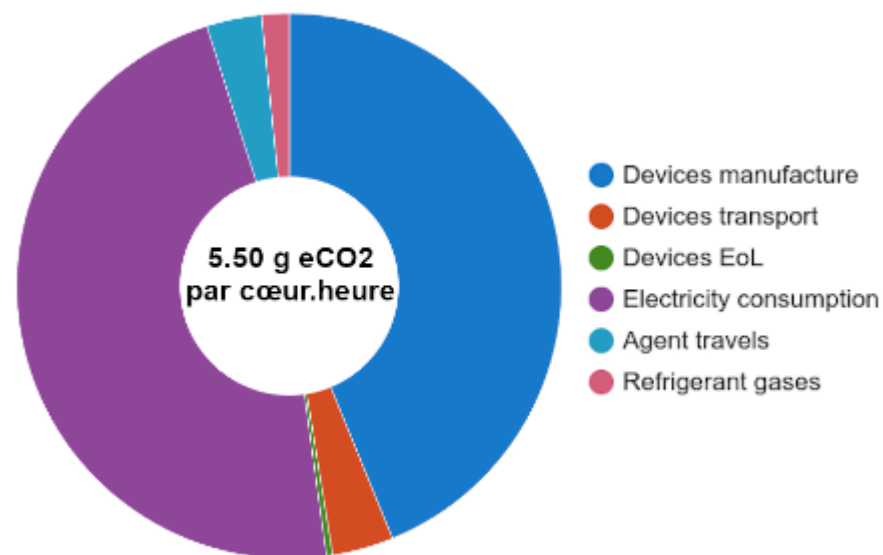


## 5. Impact carbone du calcul BGES du cluster GenoToul

- Durée de fonctionnement des appareils de **7 ans**
- Efficacité de la consommation d'énergie (PUE) du centre d'hébergement de 1,4
- Facteur d'émission de 0,1080 kg CO<sub>2</sub>e / kWh
- Temps de calcul total en 2019 de **18.000.000 heures**

# 5. Impact carbone du calcul BGES du cluster GenoToul

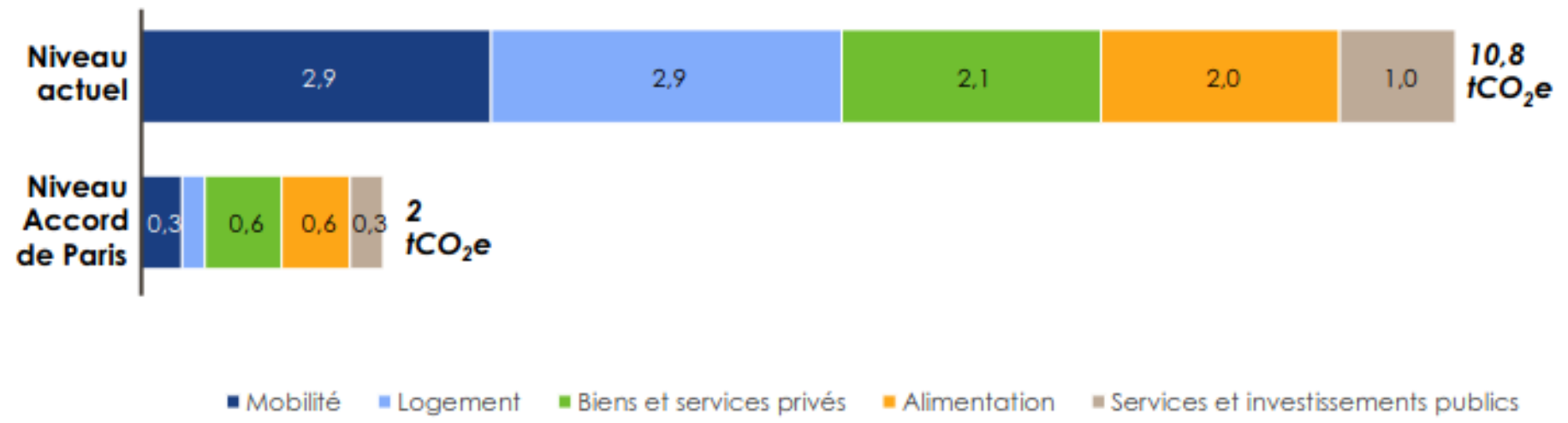
	Émissions totales de GES en 2019 (kg CO <sub>2</sub> e)	Émissions de GES par cœur.heure (g CO <sub>2</sub> e)
Fabrication du matériel	43 373	2.41
Transport du matériel	3 519	0.20
EoL du matériel	310	0.02
Consommation électrique	46 799	2.60
Transport des agents	3 250	0.18
Fluide frigorigène	1 673	0.09
<b>TOTAL</b>	<b>98 924</b>	<b>5.50</b>



# 5. Impact carbone du calcul

## BGES du cluster GenoToul

Empreinte carbone moyenne d'un Français  
tCO<sub>2</sub>



\* Source Carbone 4, 2019. Faire sa part ? Pouvoir et responsabilité des individus, des entreprises et de l'État face à l'urgence climatique.

**Budget carbone ≈ 370 000 heures CPU**

# Quizz

<https://view.genial.ly/64f5de6285e901001079880b/interactive-content-m2-metagenomic>



# TP Partie 1

<https://forgemia.inra.fr/genotoul-bioinfo/metagwgs/-/wikis/TP-metaG-HiFi-several-samples>

**A la fin : Retour d'expérience en commun : que pensez-vous de cette manière de travailler ? Quels en sont les avantages et les inconvénients ? En terme de lancement, d'accès aux résultats, de gestion des erreurs etc**

# Automatisation via NextFlow

- Permet d'enchaîner des traitements (« process ») drivés par les données (input/output) définis dans les « channels ».  
<https://www.nextflow.io/docs/latest/basic.html>
- Gère aussi de manière transparente la parallélisation sur différentes infra
- Permet de relancer après une erreur et/ou une correction. Ne relance pas ce qui s'est bien terminé et n'est pas impacté par la modification.
- La commande lancée peut être un exécutable, un script perl, python, bash ou R...
- Le script NextFlow en lui-même est en Groovy
- Grosse communauté nfCore : <https://nf-co.re/pipelines>

# Exemple de syntaxe Groovy

```
process blastThemAll {  
  input:  
  path query_file  
  
  "blastp -query ${query_file} -db nr"  
}
```

```
workflow {  
  def proteins = Channel.fromPath( '/some/path/*.fa' )  
  blastThemAll(proteins)  
}
```

Démarrera lorsque  
le channel ne sera  
pas vide



# Les répéteurs

```
process alignSequences {
  input:
  path seq
  each mode
  each path(lib)

  .....
```

t\_coffee -in \$seq -mode \$mode -lib \$lib > result

```
.....
}
```

```
workflow {
  sequences = Channel.fromPath('*fa')
  methods = ['regular', 'espresso']
  libraries = [ file('PQ001.lib'), file('PQ002.lib'), file('PQ003.lib') ]

  alignSequences(sequences, methods, libraries)
}
```

**Doc très bien faite : <https://www.nextflow.io/docs/latest/process.html>**

**Patterns : <https://github.com/nextflow-io/patterns>**

# Singularity/ Apptainer

**Un conteneur** : vous permet d'exécuter une ou plusieurs applications linux à l'intérieur d'un environnement isolé et reproductible qui ne dépend que du noyau linux de la machine sur laquelle vous êtes. Un conteneur ressemble à une machine virtuelle, sauf qu'il n'embarque pas nécessairement un système d'exploitation au complet, ce qui lui permet de se lancer en quelques secondes et d'être plus léger.

**Singularity** : L'objectif initial est de proposer une solution de conteneurisation adaptée aux besoins des scientifiques qui doivent exécuter des applications conteneurisées sur des clusters de calculs (HPC).

A la différence d'autres systèmes de conteneurs (comme Docker), Singularity ne demande aucun droit de type administrateur, aucun démon, ne virtualise pas le réseau et dialogue directement avec le système de fichiers de son hôte. Chaque conteneur est lancé et s'arrête en même temps que l'application qu'il encapsule.

# Singularity / Apptainer - Vocabulaire

**Image** : Comme dans le cadre des machines virtuelles, on appelle "image" une description statique de conteneur, une sorte de photographie de machine, que vous pouvez échanger avec vos collaborateurs, et à partir de laquelle on peut instancier et exécuter des conteneurs. Singularity possède son propre mécanisme d'images, mais peut aussi s'interfacer avec les images Docker.

## **Conteneur** :

Une machine virtuelle légère, chargée en mémoire, qui sert à lancer une application au sein d'un environnement isolé et reproductible. Un conteneur s'instancie à partir d'une image.

## **Registre** :

Entrepôt où l'on stocke des images prêtes à l'emploi. Le registre central officiel de Singularity est consultable sur la toile à l'adresse <https://singularity-hub.org/> (<https://singularity-hub.org/>).

# TP Partie 2

<https://forgemia.inra.fr/genotoul-bioinfo/metagwgs/-/wikis/TP-metaG-HiFi-several-samples>

**Nous répondrons à l'oral ensemble au 5 premières questions. Puis vous commencerez votre compte-rendu de TP individuel en répondant à l'écrit aux questions suivantes posées dans l'ordre.**

# TP Partie 3

<https://forgemia.inra.fr/genotoul-bioinfo/metagwgs/-/wikis/TP-metaG-HiFi-several-samples>

**Vous continuerez votre compte-rendu de TP individuel en répondant aux questions posées dans l'ordre.**

**Vous me l'enverrez par mail ce soir avant 17h00 :  
claire.hoede@inrae.fr**